

# Journal of Applied Psychology

Edited by

Donald G. Paterson  
University of Minnesota

---

## Consulting Editors

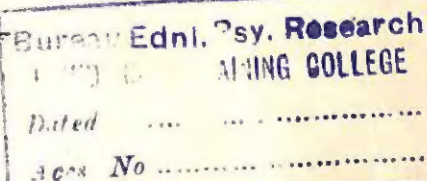
George K. Bennett, *Psychological Corporation*  
Harold E. Burtt, *Ohio State University*  
Allen L. Edwards, *University of Washington*  
Clifford E. Jurgensen, *Minneapolis Gas Co.*  
Irving Lorge, *T. C., Columbia University*  
Quinn McNemar, *Stanford University*  
Alexander Mintz, *City College of New York*

James P. Porter, *Claverack, New York*  
Harold F. Rothe, *Fairbanks, Morse and Co.,  
Beloit, Wis.*  
Julian B. Rotter, *Ohio State University*  
Edward K. Strong, Jr., *Stanford University*  
Donald E. Super, *T. C., Columbia University*  
Morris S. Viteles, *University of Pennsylvania*  
Alfred C. Welch, *Knox-Reeves, Minneapolis*



---

## Volume 38, 1954



Published Bi-monthly by the American Psychological Association, Inc.  
Prince and Lemon Sts., Lancaster, Pa.

Entered as second-class matter, August 19, 1943, at the post office at Lancaster, Pa., under the act of March 3, 1879  
Acceptance for mailing at the special rate of postage provided for in paragraph (d-2), Section 34.40,  
P. L. & R. of 1948, authorized October 10, 1947

Copyright 1954 by the American Psychological Association, Inc.

# Contents of Volume 38

## Articles

Andrews, T. G., Smith, D. D. and Kahn, L. A. An Empirical Analysis of the Effectiveness of Psychological Warfare . . . . .	240
Anikeeff, A. M. Attitudes on Social Issues of Business Administrators and Students in a School of Business Administration . . . . .	407
Anikeeff, A. M. Index of Collaboration for Test Administrators . . . . .	174
Anikeeff, A. M. Scholastic Achievement of Extension and Regular College Students . . . . .	171
Appel, V. and Kipnis, D. The Use of Levels of Confidence in Item Analysis . . . . .	256
Ash, P. Reliability and Validity of the Kopas Personnel Test Battery . . . . .	155
Balinsky, B. and Hujsa, C. Performance of College Students on a Mechanical Knowledge Test . . . . .	111
Bayton, J. A. and Thomas, C. M. Comparative and Single Stimulus Methods in Determining Taste Preferences . . . . .	443
Bendig, A. W. Reliability and the Number of Rating Scale Categories . . . . .	38
Bendig, A. W. Reliability of Short Rating Scales and the Heterogeneity of the Rated Stimuli . . . . .	167
Bendig, A. W. and Sprague, J. L. The Guilford Zimmerman Temperament Survey as a Predictor of Achievement Level and Achievement Fluctuation in Introductory Psychology . . . . .	409
Bernberg, R. E. Personality Correlates of Social Conformity . . . . .	148
Bernstein, L. An Application of Rogerian Concepts to Nurse-Patient Relationships . . . . .	324
Brayfield, A. H., Kennedy, Jr., C. E. and Kendall, W. E. Social Status of Industries . . . . .	213
Briggs, S. J., McCormick, E. J. and Kephart, N. C. The Effect of Hammer Size on Efficiency in the Task of Nailing . . . . .	1
Browne, R. C. Figure and Ground in a Two Dimensional Display . . . . .	462
Bruce, M. M. A Sales Comprehension Test . . . . .	302
Callis, R., Engram, W. C. and McGowan, J. F. Coding the Kuder Preference Record—Vocational . . . . .	359
Clark, J. G. and Owens, W. A. A Validation Study of the Worthington Personal History Blank . . . . .	85
Clark, K. E. and Gee, H. H. Selecting Items for Interest Inventory Keys . . . . .	12
Clements, F. E., Bayton, J. A. and Bell, H. P. Method of Single Stimulus Determinations of Taste Preference . . . . .	446
Comrey, A. L. and Deskin, G. Further Results on Group Manual Dexterity in Men . . . . .	116
Comrey, A. L. and Deskin, G. Group Manual Dexterity in Women . . . . .	178
Cook, W. W. and Medley, D. M. Proposed Hostility and Pharisaic-Virtue Scales for the MMPI . . . . .	414
Daniels, H. W. and Edgerton, H. A. The Development of Criteria of Safe Operation for Groups . . . . .	47
Davis, R. A. Note on Age and Productive Scholarship of a University Faculty . . . . .	318
Dingman, H. F. and Guilford, J. P. A New Method for Obtaining Weighted Composites of Ratings . . . . .	305
Di Vesta, F. J. The Effect of Methods of Presentation and Examining Conditions on Student Achievement in a Correspondence Course . . . . .	253
Di Vesta, F. J. Instructor-Centered and Student-Centered Approaches in Teaching a Human Relations Course . . . . .	329
Di Vesta, F. J. Subscore Patterns on ACE Psychological Examination Related to Educational and Occupational Differences . . . . .	248



Falk, G. H. and Bayroff, A. G. Rater and Technique Contamination in Criterion Ratings.....	100
Gaiennie, L. R. Organization Control in Business.....	289
Garvey, W. D. and Knowles, W. B. Pointing Accuracy of a Joy Stick Without Visual Feedback.....	191
Glaser, R. and Jacobs, O. Predicting Achievement in Medical School: A Comparison of Preclinical and Clinical Criteria.....	245
Gordon, D. A., Zeidner, J., Zagorski, H. J. and Uhlaner, J. E. Visual Acuity Measurements by Wall Charts and Ortho-Rater Tests.....	54
Grace, H. A. Facilitating Legislative Research.....	268
Graham, W. R. Identification and Prediction of Two Training Criterion Factors....	96
Guilford, J. P. The Validation of an "Indecision" Score for Prediction of Proficiency of Foremen.....	224
Gustad, J. W. Vocational Interests and Socio-Economic Status.....	336
Harris, S. J. and Smith, K. U. Dimensional Analysis of Motion: VII. Extent and Direction of Manipulative Movements as Factors in Defining Motions.....	126
Hay, E. N. Comparative Validities in Clerical Testing.....	299
Hendrix, O. R. "A Note" Acknowledged.....	9
Herzberg, F. Temperament Measures in Industrial Selection.....	81
Hollander, E. P. Peer Nominations on Leadership as a Predictor of the Pass-Fail Criterion in Naval Air Training.....	150
Hollander, E. P. and Bair, J. T. Attitudes Toward Authority-Figures as Correlates of Motivation Among Naval Aviation Cadets.....	21
Holmen, M. G. The Specialization Level Scale for the Strong Vocational Interest Blank.....	159
Humm, D. G. and Humm, K. A. Discussion of Gilliland and Newman's "The Humm-Wadsworth Temperament Scale as an Indicator of the 'Problem' Employee".....	131
Jacobs, R. A Note on "Predicting Success in Elementary Accounting".....	7
Jenkins, W. L. and Karr, A. C. The Use of a Joy-Stick in Making Settings on a Simulated Scope Face.....	457
Johnson, R. J. Relationship of Employee Morale to Ability to Predict Responses....	320
Kephart, N. C. and Deutsch, S. Effect of Illumination on Scores with Instrument Acuity Tests.....	59
Layton, W. L. The Relation of Ninth Grade Test Scores to Twelfth Grade Test Scores and High School Rank.....	10
Levine, P. R. and Wallen, R. Adolescent Vocational Interests and Later Occupation.....	428
Lincoln, R. S. Rate Accuracy in Handwheel Cranking.....	195
Littman, R. A. and Manning, H. M. A Methodological Study of Cigarette Brand Discrimination.....	185
Lockman, R. F. Some Relationships Between the MMPI and a Problem Checklist.....	264
Longstaff, H. P. Practice Effects on the Minnesota Vocational Test for Clerical Workers.....	18
MacKinney, A. C. and Jenkins, J. J. Readability of Employee's Letters in Relation to Occupational Level.....	26
MacLean, A. G. and Tait, A. T. Some Computational Short-Cuts in the Development or Analysis of Tests.....	260
MacPhail, A. H. Interest Patterns for Certain Degree Groups on the Lee-Thorpe Occupational Interest Inventory.....	164
Maloney, P. W. Comparability of Personal Attitude Scale Administration with Mail Administration with and without Incentive.....	238



Marchetti, P. V. Manager-Employee "Understanding" in the Retail Grocery and Meat Market.....	216
Mason, H. M. A Comparative Evaluation of Two Approaches to Job-Knowledge Test Construction.....	384
McArthur, C. Long-Term Validity of the Strong Interest Test in Two Subcultures..	346
McCormick, E. J. and North, W. E. The Analysis of an Experimental Job Evaluation System as Applied to Enlisted Naval Jobs.....	233
McQuitty, L. L., Wrigley, C. and Gaier, E. L. An Approach to Isolating Dimensions of Job Success.....	227
Meyer, H. D. and Pressel, G. L. Personality Test Scores in the Management Hierarchy.....	73
Minnesota State Employment Service in Cooperation with the U. S. Employment Service, U. S. Department of Labor, Washington, D. C. Standardization of the GATB for the Occupation of Tabulating Machine Operator.....	297
Mintz, A. The Inference of Accident Liability from the Accident Record.....	41
Mintz, A. Time Intervals Between Accidents.....	401
Mosel, J. N. Response Reliability of the Activity Vector Analysis.....	157
Muckler, F. A. and Matheny, W. G. Transfer of Training in Tracking as a Function of Control Friction.....	364
Neidt, C. O. and Malloy, J. P. A Technique for Keying Items of an Inventory to Be Added to an Existing Test Battery.....	308
Newman, S. H. Quantitative Analysis of Verbal Evaluations.....	293
Owens, Jr., W. A. The Retest Consistency of Army Alpha After Thirty Years.....	154
Owens, Jr., W. A. A Reply to Drs. Peck-Stephenson.....	371
Patton, Jr., W. M. Studies in Industrial Empathy: III. A Study of Supervisory Empathy in the Textile Industry.....	285
Peck, R. F. and Stephenson, W. A Correction of the Clark-Owens Validation Study of the Worthington Personal History Technique.....	368
Powers, R. D. Sampling Problems in Studies of Writing Style.....	105
Rosenberg, N. and Izard, C. E. Vocational Interests of Naval Aviation Cadets.....	354
Ross, S., Hussman, T. A. and Andrews, T. G. Effects of Fatigue and Anxiety on Certain Psychomotor and Visual Functions.....	119
Rust, R. M. and Ryan, F. J. The Strong Vocational Interest Blank and College Achievement.....	341
Scales, E. M. and Chapanis, A. The Effect on Performance of Tilting the Toll-Operator's Keyset.....	452
Siegel, A. I. The Check List as a Criterion of Proficiency.....	93
Siegel, A. I. An Experimental Evaluation of the Sensitivity of the Empathy Test....	222
Siegel, A. I. Retest-Reliability by a Movie Technique of Test Administrators' Judgments of Performance in Process.....	390
Singer, S. L. and Steffire, B. The Relationship of Job Values and Desires to Vocational Aspirations of Adolescents.....	419
Spector, A. J. Influences on Merit Ratings.....	393
Stone, J. B. Differential Prediction of Academic Success at Brigham Young University.....	109
Stordahl, K. E. Permanence of Interests and Interest Maturity.....	339
Stordahl, K. E. Permanence of Strong Vocational Interest Blank Scores.....	423
Strong, Jr., E. K. Validity versus Reliability.....	103
Teel, K. S. and Du Bois, P. H. Psychological Research on Accidents: Some Methodological Considerations.....	397
Tiffin, J. and Winick, D. M. A Comparison of Two Methods of Measuring the Attention-Drawing Power of Magazine Advertisements.....	272



Tinker, M. A. Readability of Mathematical Tables.....	436
Van Zelst, R. H. The Effect of Age and Experience upon Accident Rate.....	313
Van Zelst, R. H. and Kerr, W. A. Personality Self-Assessment of Scientific and Technical Personnel.....	145
Washburne, N. F. and Andrew, D. C. Relation of Scholastic Aptitude to Socio-economic Status and to a Rural-to-Urban Continuum.....	113
Wexner, L. B. The Degree to Which Colors (Hues) Are Association with Mood-Tones.....	432
Wilson, R. C. and Comrey, A. L. A Short Method of Factor Analysis.....	181
Wilson, R. C., High, W. S., Beem, H. P. and Comrey, A. L. A Factor-Analytic Study of Supervisory and Group Behavior.....	89
Witryol, S. L. Scaling Procedures Based on the Method of Paired Comparisons....	31
Wood, T. L. The Relationship Between Mechanical Aptitude and Proficiency Tests for Air Force Mechanics.....	381

### Book Reviews

Anonymous. Army Personnel Tests and Measurements: Harold E. Burt.....	280
Berdie's Roles and Relationships in Counseling: Arthur H. Brayfield.....	375
Bross' Design for Decision: Allen L. Edwards.....	376
Bullock's Social Factors Related to Job Satisfaction, a Technique for the Measurement of Job Satisfaction: Howard L. Roy.....	142
Buros' The Fourth Mental Measurements Yearbook: Charles N. Morris.....	281
Coombs' A Theory of Psychological Scaling: Marvin D. Dunnette.....	66
Husband's The Psychology of Successful Selling: Brent Baxter.....	65
Illuminating Engineering Society's Recommended Practice for Residence Lighting: Miles A. Tinker.....	141
Jahoda, Deutsch, and Cook's Research Methods in Social Relations, with Especial Reference to Prejudice; Vol. I: Basic Processes; Vol. II: Selected Techniques: Harrison G. Gough.....	66
Jennings' Techniques of Successful Foremanship: Theodore R. Lindbom.....	207
Kinsey, Pomeroy, Martin, Gebhard, <i>et al.</i> Sexual Behavior in the Human Female; Hiltner's Sex Ethics and the Kinsey Reports; and Aberle and Corner's Twenty-five Years of Sex. Research History of the National Research Council Committee: Donald G. Paterson.....	205
Lawshe's Psychology of Industrial Relations: A. S. Thompson.....	64
Leitner's Hypnotism for Professionals: William T. Heron.....	207
Lincoln's Incentive Management: Albert S. Thompson.....	135
Lundin's An Objective Psychology of Music: Kate Hevner Mueller.....	206
Marketing and Social Research Division of the Psychological Corporation's The Measured Effectiveness of Employee Publications: Donald G. Paterson.....	279
Comment on Preceding Review: Charles L. Vaughn.....	280
McFarland's Human Factors in Air Transportation: George K. Bennett.....	65
Montagu's The Natural Superiority of Women: Leona E. Tyler.....	208
New York Academy of Medicine and the Josiah Macy, Jr. Foundation's (Transactions of the Conference on) Morale—and the Prevention and Control of Panic: Clark L. Hosmer.....	67
Personality: Symposia on Topical Issues, Vol. 1, Nos. 3 and 4 (pp. 213-388): Frank A. Pattie.....	70
Powers and Witmer's An Experiment in the Prevention of Delinquency: The Cambridge-Somerville Youth Study: Elio D. Monachesi.....	68
Redfield's Communication in Management: George Klare.....	138



Remmers' Introduction to Opinion and Attitude Measurement: Sidney S. Goldish..	377
Schlotter and Svendsen's An Experiment in Recreation with the Mentally Retarded: Harriet E. Blodgett.....	379
Sherif and Wilson's Group Relations at the Crossroads: Bernard M. Bass.....	378
Traxler, Jacobs, Selover, and Townsend's Introduction to Testing and the Use of Test Results in Public Schools: Marjorie Olsen.....	68
Tuckman and Lorge's Retirement and the Industrial Worker: Prospect and Reality: Marvin D. Dunnette.....	375
Tyler's The Work of the Counselor: Donald E. Super.....	139
Viteles' Motivation and Morale in Industry: Clifford E. Jurgensen.....	136
Woolf and Woolf's The Student Personnel Program: John W. Gustad.....	206

### Applied Psychology in Action

Colmen, J. G. Psychological Research in Personnel Administration.....	61
Dahlstrom, W. G. Personnel Psychology and Small Business.....	203
Dvorak, B. J. GATB in Foreign Countries.....	373
Epstein, M. A Note on "The Non-Directive Approach in Advertising Appeals".....	133
Hadley, H. D. Reply to Dr. Wells and to Miss Epstein.....	202
Jurgensen, C. E. Reporting Employment Test Scores to Supervisors.....	277
Kerr, W. The Measurement of Academic Freedom.....	134
Knauff, E. B. Time Limit versus Work Limit Methods of Test Administration...	62
Murrell, K. F. H. Note on the Work of the British Standards Institution.....	202
Wells, F. L. Comment on Word Meaning.....	133
Employee Opinion Surveys.....	63
Legal Status of Advertising and Marketing Psychology Experts.....	276

### Miscellaneous

New Books, Monographs, and Pamphlets.....	71, 143, 210, 282, 380, 468
---	-----------------------------



# Journal of Applied Psychology

VOL. 38, No. 1

FEBRUARY, 1954

## The Effect of Hammer Size on Efficiency in the Task of Nailing \*

Stewart J. Briggs, E. J. McCormick, and N. C. Kephart

*Occupational Research Center, Purdue University*

Any hardware store salesman "knows" what size and type of hammer to use with different sizes and types of nails. On the basis of the intuitive knowledge that impregnates the atmosphere of any hardware store, the salesman will sell to the home craftsman a small hammer to use with small nails and a large hammer to use with larger nails. There has apparently never been any empirical evidence, however, to verify or deny the salesman's judgment on these matters. This study was designed to provide at least a fragment of such empirical evidence.

More specifically the investigation was carried out to determine the relationships, in terms of efficiency in nailing, between sizes and types of hammers and sizes and types of nails as used by home craftsmen. Six hammers were used in the experiment, four of them being claw hammers and two rip hammers. Five sizes of finishing nails and five sizes of common nails were used.

### Experimental Procedures

While it would have been desirable to establish conditions that simulated those which the home craftsman would meet, it was not possible to accomplish this objective entirely because of the need to exercise experimental controls.

*Pilot Study.* A pilot study was carried out with one subject. On the basis of the pilot study, certain observations were made and these were used in developing the procedures for the experiment proper. Following are the observations that resulted from the pilot study:

\* Appreciation is expressed to Mr. L. A. O'Connor, Store Manager, and Mr. Myron Burkenpas, Manager of the Hardware Department, Sears Roebuck and Company, Lafayette, Indiana, for the loan of the hammers for this experiment.

1. The measured time of the task was a more suitable criterion of performance than number of strikes of hammer since it takes into account the effect of bent nails.

2. The wood used should be of uniform grain and of medium hardness.

3. The optimum number of nails to be driven for each nail-hammer combination was about three.

4. Rest periods were necessary to reduce variance due to fatigue.

*Subjects.* Six subjects were used in the experiment. All of the subjects selected had had experience as home craftsmen, yet not as professional carpenters. The subjects were all males between the ages of 21 and 39 years, and were associated with Purdue University; one was a professor of psychology, four were graduate students in psychology, and one was an undergraduate in the field of engineering.

*Materials.* The following materials were used:

1. Six hammers were used: four were classified commercially as 7, 10, 13, and 16 oz. claw hammers; and two were 16 and 20 oz. rip hammers. The hammers were marked with letters for identification. It should be noted that weight size refers to the weight of the hammer head, and the terms "claw" and "rip" refer to the shape of the head.

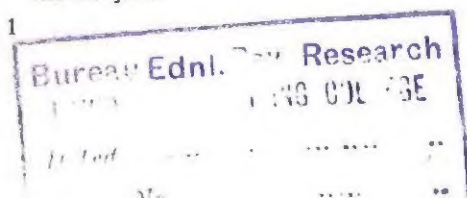
2. Nails of the following types were used: 4, 6, 8, 10, and 16 penny common wire nails, and 2, 4, 6, 8, and 10 penny wire finishing nails. The nails varied in length by half inch intervals and increased in gauge with the larger sizes. The finishing nails were of smaller gauge than their penny equivalents in common nails.

3. One eight foot top grade fir 2 x 4 per subject.

4. Two sawhorses approximately 34 inches in height equipped with a wooden groove to hold the 2 x 4 in place during the experiment.

5. Nail containers: one wooden nail bin of nine compartments and one can for holding the largest nails. Each container was marked with the size of the nail it contained.

6. One table on which the nail bins were placed, positioned to place the nails within easy reach of the subject.





7. One stop watch calibrated in hundredths of a minute.

*Warm-up Period.* Each subject was allowed a short warm-up period during which he drove up to a total of ten nails of various sizes using three or four different hammers. This warm-up period ended when the subject said he was ready to begin the experiment.

*Experimental Sequence.* Each subject drove nails of each of the ten types and sizes with each of the six hammers, making a total of sixty combinations of nails and hammers for each subject. A deck of sixty IBM cards was prepared for each subject, each card representing a combination of one nail type and size and one hammer. These cards were thoroughly shuffled, and the subject followed the randomized order that resulted from this shuffling; as the subject completed any one combination of nail and hammer, the experimenter would tell him what combination to use next. Three nails of each type and size were driven by the subject. The experimenter timed the subject on the total time required to drive the three nails from the time the subject grasped the first nail until he had completed driving the third one. The time records were recorded on the cards. (The time records were later punched into these cards for use in statistical analysis.)

Short rest pauses of approximately one-half minute were introduced between each of the sixty combinations. Two longer rest periods of ten minutes divided the experimental time into roughly three equal intervals.

*Instructions to the Subject.* The following instructions were given to the subject:

"The task involves driving nails into this  $2 \times 4$ . You are to drive the nails in sets of three; that is, I will measure the time from the instant you grasp the first nail until you have finished driving the third. Drive each nail until its head is flush with the board before driving the next. Try not to mar the wood. If the nail starts to bend, try to correct it; and if it seems too bent, pull it out and use another nail in its place.

"Before each set I will tell you which hammer and nail to use. These are identified by the letters on the hammer and the numbers on the compartments of the nail bins. You will then select the proper hammer and hold it in the hand you wish to hammer with. When you have located the proper nail bin, say *ready* after which I will say *go*. Then grasp the nail and start hammering. Drive the nails as fast as possible, remembering that bent nails will slow you up. Are there any questions?"

## Results

The data were treated statistically using an analysis of variance to identify significant variables and interactions. The data were

Table 1  
Analysis of Variance

Source	Mean Square	d.f.	F
Subjects	155.1	5	
Hammers	734.3	5	16.9**
Nails	2344.4	9	50.3**
Hammers×Subjects	43.4	25	1.08
Nails×Subjects	46.7	45	1.16
Hammers×Nails	73.0	45	1.82**
Hammers×Nails×Subjects	40.2	225	
Total		359	

Special Analyses			
Source	Mean Square	d.f.	F
Hammers			
16 oz. claw—16 oz. rip	6.07	1	.14
Nails (4, 6, 8, 10 penny)			
Finishing—Common	417.9	1	8.94**

\*\* Denotes significance at the 1% level of confidence.

further treated using the process described by Tukey<sup>1</sup> to break up the data into significantly different groups.

*Analysis of Variance.* The results of the analysis of variance are presented in Table 1. These findings may be interpreted as showing that the variance within the different sizes of hammers as well as different sizes of nails is statistically significant. Only one claw hammer and one rip hammer were of comparable size (16 oz.) and they were found not to be significantly different. There was a significant difference between the two types of nails (common and finishing) when only those sizes represented in both types (4, 6, 8, 10 penny) were considered. The finishing nails were driven more slowly than were their penny equivalents in common nails.

There was no significant interaction either between hammers and subjects or between nails and subjects. There was found a significant variance ratio in the hammer by nail interaction. That is, certain hammer and nail combinations can be considered better than others when driving time is used as a criterion. To locate these specific combinations, the Tukey process was used.

<sup>1</sup> Tukey, J. W. Comparing individual means in the analysis of variance. *Biometrics*, 1949, 5, 99-114.



Table 2  
Hammer Size Groups for Specific Nails

Nail	Hammer	Mean Time	Sub-group	Nail	Hammer	Mean Time	Sub-group
4 C	16 Claw	21.17	*	2 F	16 Rip	22.67	*
4 C	20 Rip	23.17	*	2 F	13 Claw	26.00	*
4 C	13 Claw	23.50	*	2 F	10 Claw	26.00	*
4 C	16 Rip	23.67	*	2 F	7 Claw	26.33	*
4 C	7 Claw	25.17	*	2 F	16 Claw	29.50	*
4 C	10 Claw	25.33	*	2 F	20 Rip	34.00	*
6 C	16 Claw	25.33	*	4 F	16 Rip	24.17	*
6 C	20 Rip	26.83	*	4 F	20 Rip	27.00	*
6 C	16 Rip	26.83	*	4 F	16 Claw	28.00	*
6 C	13 Claw	27.50	*	4 F	13 Claw	28.67	*
6 C	10 Claw	31.83	*	4 F	10 Claw	29.17	*
6 C	7 Claw	33.00	*	4 F	7 Claw	30.00	*
8 C	13 Claw	30.67	*	6 F	20 Rip	28.33	*
8 C	16 Claw	31.17	*	6 F	16 Claw	28.67	*
8 C	20 Rip	31.33	*	6 F	16 Rip	29.83	*
8 C	16 Rip	34.00	*	6 F	13 Claw	30.33	*
8 C	10 Claw	36.33	*	6 F	7 Claw	33.50	*
8 C	7 Claw	40.17	*	6 F	10 Claw	35.50	*
10 C	16 Claw	34.00	I	8 F	16 Rip	33.17	*
10 C	20 Rip	34.17	I	8 F	20 Rip	33.33	*
10 C	16 Rip	36.67	I	8 F	16 Claw	33.67	*
10 C	13 Claw	37.00	I	8 F	13 Claw	34.83	*
10 C	10 Claw	45.17	II	8 F	7 Claw	39.33	*
10 C	7 Claw	52.50	II	8 F	10 Claw	40.83	*
16 C	20 Rip	41.50	I	10 F	16 Rip	36.17	I
16 C	16 Rip	46.17	I	10 F	13 Claw	37.33	I
16 C	13 Claw	47.00	I	10 F	16 Claw	38.00	I
16 C	16 Claw	47.17	I	10 F	20 Rip	40.17	I
16 C	10 Claw	54.33	I	10 F	10 Claw	43.67	I
16 C	7 Claw	66.67	II	10 F	7 Claw	51.83	II

## Legend:

## Nail Type:

Number refers to penny size (2 = 2 penny, 4 = 4 penny, etc.).

C = Common wire nail; F = Finishing wire nail.

## Hammers:

Number refers to size (16 = 16 oz. hammer, etc.).

## Subgroups:

\* = No significantly different subgroups formed (at 1% level).

I = Subgroup with significantly faster driving time (at 1% level).

II = Subgroup with significantly slower driving time (at 1% level).

*Tukey Process.*<sup>2</sup> The data presented in Tables 2 and 3 represent the results of employing the technique developed by Tukey for dividing a group of means into significantly different subgroups. Table 2 shows for each nail size the subgroups that occurred between the means of the different

hammers, i.e., given a nail of a certain type and size, which hammer or hammers are the best? Table 3 shows the subgroups occurring between nails when each hammer was used. As the number of subgroups formed is dependent upon the variance of the whole group, the number of subgroups is not constant.

<sup>2</sup> Tukey, J. W. *Op. cit.*



In Tables 2 and 3 the significantly different subgroups are identified with Roman numerals. With two subgroups numerals I and II are used; with three subgroups, numerals I, II, and III. An asterisk (\*) is used where no subgroups were found.

On the basis of the subgroups that were found, it is possible to make certain general recommendations with regard to hammer-nail combinations. In Table 2 it will be noted that for three nail sizes significantly different subgroups of hammers were formed; in each

Table 3  
Significantly Different Hammer Subgroups for Specific Nails

Hammer	Nail	Mean Time	Sub-group	Hammer	Nail	Mean Time	Sub-group
7 Claw	4 C	25.17	I	16 Claw	4 C	21.17	I
7 Claw	2 F	26.33	I	16 Claw	6 C	26.50	II
7 Claw	4 F	30.00	I	16 Claw	4 F	28.00	II
7 Claw	6 C	33.00	I	16 Claw	6 F	28.67	II
7 Claw	6 F	33.50	I	16 Claw	2 F	29.50	II
7 Claw	8 F	39.33	I	16 Claw	8 C	31.17	II
7 Claw	8 C	40.17	I	16 Claw	8 F	33.67	II
7 Claw	10 F	51.83	II	16 Claw	10 C	34.00	II
7 Claw	10 C	52.50	II	16 Claw	10 F	38.00	II
7 Claw	16 C	66.67	III	16 Claw	16 C	47.17	III
10 Claw	4 C	25.33	I	16 Rip	2 F	22.67	I
10 Claw	2 F	26.00	I	16 Rip	4 C	23.67	I
10 Claw	4 F	29.17	I	16 Rip	4 F	24.17	I
10 Claw	6 C	31.83	I	16 Rip	6 C	26.83	I
10 Claw	6 F	35.50	I	16 Rip	6 F	29.83	I
10 Claw	8 C	36.33	I	16 Rip	8 F	33.17	I
10 Claw	8 F	40.83	II	16 Rip	8 C	34.00	I
10 Claw	10 F	43.67	II	16 Rip	10 F	36.17	I
10 Claw	10 C	45.17	II	16 Rip	10 C	36.67	I
10 Claw	16 C	54.33	III	16 Rip	16 C	46.17	II
13 Claw	4 C	23.50	I	20 Rip	4 C	23.17	I
13 Claw	2 F	26.00	I	20 Rip	6 C	26.83	I
13 Claw	6 C	27.50	I	20 Rip	4 F	27.00	I
13 Claw	4 F	28.67	I	20 Rip	6 F	28.33	I
13 Claw	6 F	30.33	I	20 Rip	8 C	31.33	I
13 Claw	8 C	30.67	I	20 Rip	8 F	33.33	I
13 Claw	8 F	34.83	I	20 Rip	2 F	34.00	I
13 Claw	10 C	37.00	I	20 Rip	10 C	34.17	I
13 Claw	10 F	37.33	I	20 Rip	10 F	40.17	II
13 Claw	16 C	47.00	III	20 Rip	16 C	41.50	II

Legend:

Nail Type:

Number refers to penny size (2 = 2 penny, 4 = 4 penny, etc.).

C = Common wire nail; F = Finishing wire nail.

Hammer:

Number refers to size (16 = 16 oz. hammer, etc.).

Subgroups:

I = Subgroup with significantly faster driving time (at 1% level).

II = Subgroup significantly slower than I (at 1% level).

III = Subgroup significantly slower than I and II (at 1% level).



such case the hammers in subgroup I are to be recommended over those in subgroup II.

From an overview of Table 2 it may be concluded that the 10 and 7 oz. hammers would not be good general purpose hammers under conditions similar to those in this experiment. The 16 oz. and 20 oz. rip hammers, and the 13 and 16 oz. claw hammers appear to have better all-around characteristics.

It will be observed in Table 3 (which deals with subgroups of nails for individual hammers) that significant subgroups of nails were formed for each hammer. In some cases two subgroups were formed; in such cases the nails included in subgroup I were driven significantly faster with the hammer in question than those nails in subgroup II. In the case of other hammers, three subgroups of nails were formed; in such cases the nails included in subgroup II were driven faster than those in subgroup III, and those in subgroup I were driven faster than those in II. A general observation of Table 3 would suggest that smaller nails were driven faster than larger nails, which of course is to be expected. It might be noted that the 4 penny common nails were in subgroup I for all hammers.

It should be stressed that the data in Table 2 are applicable for the situation where nails of a given size and type are to be driven, and it is desired to select the most efficient available hammer for the job; this covers most home craftsman situations. However, it is conceivable that situations would arise where various nails could be used equally well, where they are to be used in quantities, and possibly where there is a limited choice of hammers; in such a situation the data in Table 3 would be appropriate since they show the relative speeds with which various nails were driven with specified hammers.

#### Discussion

The results are not entirely consistent with the salesman's intuitive judgment. The large hammers were found to be better with the larger nails; however, for smaller nails, the smaller hammers were *not* significantly better. This may be a function of the range of

nail sizes; if small brads had been included, the smaller hammers might have been found to be better, although it should be noted that two penny finishing nails (which were used in the experiment) are quite small, being only one inch long and of small gauge.

The two 16 oz. hammers were expected to have the same hammering characteristics as they differ only slightly in the shape of the nail pulling part of the hammer head. This small deviation would not be expected to affect the balance of the hammer seriously, and no significant differences were found between these two types of hammers.

The statistically significant difference between common and finishing nails was not entirely expected. It was thought at first that the greater diameter of the common nails would offer greater resistance and hence slow up the driving. However, this same greater diameter presumably tended to reduce the time lost due to nail bending. It is also possible that the appearance to the subjects of greater frailty of the finishing nails may have made them somewhat more cautious (and therefore slower) in driving the finishing nails.

It should be kept in mind that the experiment was conducted using only fir. While this is a commonly used wood by the home craftsman, the results cannot be generalized with assurance to harder or softer woods. It might be hypothesized that the results are more general than this experiment indicates, as the relationship of the weight of the hammer to the bending resistance of the nail might be more crucial than the hardness of the wood. If this were true, it would indicate that the skill of the hammerer is most important in nailing into harder woods; but the relationship of the hammer and nail would be the same. Further research would, of course, be required to explore such variables.

It is recognized that time is not necessarily the best criterion of performance for every situation; for instance, in cabinet work or finish carpentry, lack of marks in the wood undoubtedly would be a better criterion of performance than speed. It should be noted that in this experiment an attempt was made through instructions and reminders to con-



trol the marring of the wood, but this was not completely successful.

The entire field of study of the tools of the home craftsman is lacking in systematic investigation. The methods of analysis of variance and Tukey's process appear to be powerful tools for study in this area because they allow more than one variable to be studied at a time and yet permit specific recommendations to be made. In studies of this field, it seems advisable to plan a pilot experiment (as the one carried out in this study) to locate and control some of the unexpected experimental difficulties so they will not interfere with the main study.

#### Summary

This study was carried out to determine the relationship in terms of efficiency of use between six hammers (7, 10, 13, and 16 oz. claw and 16 and 20 oz. rip hammers) and ten nails (4, 6, 8, 10, and 16 penny common and 2, 4, 6, 8, and 10 penny finishing nails) when used by home craftsmen. The six subjects were home craftsmen without professional carpentering experience. The subjects drove a set of three nails into a fir  $2 \times 4$  for

each of the sixty possible hammer and nail combinations. Time was the criterion of performance.

Analysis of variance was used on the data, and the results indicated:

1. The variance in time due to the different hammers was statistically significant.
2. The variance in time due to the different nails was statistically significant.
3. There was no statistically determined difference between the 16 oz. rip and claw hammers.
4. The finishing nails were slower to drive than the common nails.
5. The variance in time due to nail by hammer interaction was significant.

The data were further treated by Tukey's process to locate various significant subgroups of hammer-nail combinations. Specific recommendations were made considering first the hammer, then the nail, as the independent variable.

The methods used were felt to be applicable to other research in the field of home craftsman's tools.

*Received April 23, 1953.*



## A Note on "Predicting Success in Elementary Accounting"

Robert Jacobs

*Educational Records Bureau, New York, N. Y.*

A study reported by O. R. Hendrix in the April, 1953 issue of *J. appl. Psychol.*<sup>1</sup> compared the validities of the 1947 Edition of the ACE Psychological Examination, the Ohio State University Psychological Test (Form 23), and the latest form (Form C) of the Orientation Test used in the accounting testing program sponsored by the American Institute of Accountants. The criterion for validity was grades received in elementary accounting by 76 men and 19 women in the College of Commerce and Industry of the University of Wyoming.

The correlations reported in this study between accounting grades and the scores on the different tests ran somewhat higher for the ACE Psychological Examination and the OSU Psychological Test than for the AIA Orientation Test.

On the basis of the data which he obtained, Hendrix concluded that "If a single test is to be utilized in predicting grades in elementary accounting, ACE Psychological Examination and OSU Psychological Test are preferable to the AIA Orientation Test." The author of the study points out that the investigation was restricted to the relationship between the test scores considered and grades and that a different pattern of relationship might be found if the criterion of validity were success in actual professional employment as an accountant.

However, the relative superiority of the AIA Orientation Test when compared with tests of general scholastic ability in predicting success in *accounting study* is a matter of concern to counselors and to teachers when the Orientation Test is used in the College Accounting Testing Program.

A considerable amount of research data has been accumulated at the project office relating to the reliability and validity of the tests used in the College Accounting Testing Pro-

gram. Some of these data are the result of research carried out at the project office; some are results of independent studies carried out by participating schools and reported to the project office. As with any program which reaches into many institutions and many different kinds of situations, the data show a rather wide range of results. In some schools, correlations between Orientation Test scores and accounting grades have been unusually high, while with other groups in different schools relationships shown have been on the disappointing side. The usual procedure in dealing with such an accumulation of data is to generalize on the basis of central tendencies of results. This procedure has been followed in reporting on the validity and the reliability of the instruments used in the College Accounting Testing Program. The point is that it is usually an unsafe procedure to generalize from a single study based on a particular group of students. If the results obtained with one group are borne out with data from similar research based on different groups, it may be safe to generalize a finding or a trend.

Most of the *comparative* validity data gathered at the project office has been concerned with a comparison of the Orientation Test and the ACE Psychological Examination. This is true because the ACE test is the most widely used test of scholastic ability at the college level, and hence, most of the questions concerning superiority of the Orientation Test coming from institutions participating in the program related to the ACE test which, commonly, was part of the battery of tests already used in the college. The data from several of these studies are shown in Table 1, together with the results reported by Hendrix.

The Table 1 correlations show varying results, but only in the Hendrix study does the difference between the pair of correlations favor the ACE test. The data reported for

<sup>1</sup> O. R. Hendrix. Predicting success in elementary accounting. *J. appl. Psychol.*, 1953, 37, 75-77.



Table 1

Correlations between Orientation Test Total Score and Grades in Accounting Courses Compared with Correlations between ACE Psychological Examination Total Score and Accounting Grades \*

Source of Data	Institution	N	$r$ Orient. vs. Grades	N	$r$ ACE vs. Grades
Project Office Study	Drake University	363	.39	294	.27
Project Office Study	U. of Louisville Group A	166	.43	161	.22
Project Office Study	U. of Louisville Group B	133	.38	134	.15
Project Office Study	Wayne University	265	.37	99	.28
Roth Study	CCNY	148	.23	148	.19
Hendrix Study	University of Wyoming	95	.32	95	.36

\* In most instances the ACE Psychological Examination is administered at the beginning of the freshman year and the Orientation Test at the beginning of the sophomore year, but before the students have completed a semester of accounting study.

the project office studies show differing N's for the comparative correlations. This may raise some questions regarding the validity of comparisons. The correlations between other test scores and accounting grades were obtained as supplementary studies following the checks on relationships between Orientation Test scores and grades. Scores on other tests were not available for all students taking the Orientation Test, with the exception of the University of Louisville Group B, but so far as is known, no bias occurred in the use of the smaller population. The difference in N is of most concern in the Drake University and Wayne University data, and it will be noted that the superiority of the Orientation Test is less noticeable in these two instances than in the case of the two University of Louisville groups where the N's are in closer agreement. Furthermore, the study in which the N's were the same, the one carried on by Roth at CCNY (unpublished), shows as much difference in favor of the Orientation Test as does Hendrix's study in favor of the ACE.

However, the point of this short note is not so much to argue the superiority of the Ori-

entation Test as to suggest the danger in generalizing the superiority of one testing instrument over another on the basis of a study using the results from a single institution.

The Hendrix study reports the only comparison between the Orientation Test and the OSU Psychological Test which has come to the attention of the project office (OSU test vs. grades = .37; AIA test vs. grades = .32). As indicated with the ACE exam data, however, it would be hazardous to generalize on the basis of this one bit of evidence.

It is believed by this writer that a further note of caution could be added to Hendrix's summary to the effect that "It does not necessarily follow that the same relationship would be obtained in a different institution and with a different group of students." The data shown in Table 1 indicate that results do differ from one group to another, and they suggest further that the general trend in comparative validity tends to favor the Orientation Test rather than the ACE Psychological Examination.

Received September 24, 1953.

Published out-of-turn by the editor.

## "A Note" Acknowledged

O. R. Hendrix

*Office of Student Personnel and Guidance, The University of Wyoming*

In the preceding article, Jacobs has suggested that a further note of caution could be added to the summary of the study reported by this writer in the April, 1953 issue of the *Journal of Applied Psychology*. The suggested note is: "It does not necessarily follow that the same relationship would be obtained in a different institution and with a different group of students."

One could certainly have no objection to such a statement. As Jacobs points out, numerous correlation studies have verified its accuracy. This writer most certainly did not intend to imply that the results of his limited study had general application. In fact, he tried to guard against such an assumption by stating in his opening paragraph that the study reported represented an investigation of the relative validity of the several tests "for predicting success in elementary accounting at the University of Wyoming" (italics added).

While concurring with Jacobs' desire to guard against generalization on the basis of a single study, one should be equally careful to keep a number of limitations of Jacobs' own study in mind while considering his statement concerning "the relative superiority of the AIA Orientation Test when compared with tests of general scholastic ability in predicting success in accounting study. . . ."

First, there is the limitation growing out of the differences in the number of cases used in the computation of the coefficients of correlation between grades and Orientation Test scores and the number of cases used in computing the coefficients of correlation between grades and the ACE. In the instance of the Wayne University data, 265 cases were used for one computation and only 99 for the second computation and in the Drake University study the number of cases was 363 in one instance and 294 in another. While the author assumed that no bias occurred in re-

ducing *N*, the possibility that bias *did* occur cannot be ruled out.

A second limitation grows out of the fact that "in most instances" the Orientation Test was administered a year later than the ACE. One might assume that learning which took place during this year had no effect on the correlation between Orientation Test scores and accounting grades. One would also have to consider the possibility that the year of learning influenced either or both test scores and grades and consequently affected the correlation between the two. If the year of learning involved any accounting, one is confronted with an interesting effort to predict aptitude for learning that which has already been learned.

A third limitation is that inherent in making generalizations about "the relative superiority of the AIA Orientation Test when compared with tests of general scholastic ability . . ." on the basis of studies limited to comparison of the Orientation Test and a *single* scholastic ability test, namely the ACE.

Possibly in an effort to keep his note brief, Jacobs has failed to mention the possibilities for more accurate prediction through the use of a number of predictors rather than a single predictor. Well-trained counselors seldom depend upon a single predictor. An increasing number of counseling agencies are constructing prediction equations based upon multiple variables. The question of whether the Orientation Test contributes significantly to such equations still has to be answered.

It is entirely possible that studies in which the above listed limitations are not operative would provide proof that the AIA Orientation Test is superior to tests of general scholastic ability. Until such studies are cited, one would seem justified in retaining an open mind on the subject.

*Received November 4, 1953.*

*Published out-of-turn by the editor.*



# The Relation of Ninth Grade Test Scores to Twelfth Grade Test Scores and High School Rank

Wilbur L. Layton

*Student Counseling Bureau, University of Minnesota*

The 9th grade is a crucial one for most students, for they, their school counselors, teachers and administrators must make decisions which are important for the students' high school careers and in fact for their entire futures. High school guidance workers test many 9th grade students in order to assist them to select appropriate high school curricula.

This study was an attempt to determine the meaning of 9th grade tests as predictors of over-all high school achievement and 12th grade test scores.

In January and February of 1949, the 1947 High School Edition of the ACE Psychological Examination was administered to approximately 15,000 ninth grade students in Minnesota through the state-wide high school testing program administered by the Student Counseling Bureau of the University of Minnesota. The students tested in this program were from schools volunteering to participate in the program at their own expense. These schools consisted of approximately 50 per cent of non-metropolitan high schools in Minnesota. Approximately 10,000 ninth graders were also given the Cooperative English Test, Form Y, Lower Level, Single Booklet Edition, Mechanics of Expression, Effectiveness of Expression and Reading Comprehension. Three

Table 1

N's, Means and Standard Deviations for 9th Grade Test Scores and 12th Grade Test Scores and High School Percentile Rank

	N	Mean	Standard Deviation
9th ACE	2,173	67.9	18.4
12th ACE	2,185	94.7	24.0
9th English	690	155.6	64.4
12th English	2,185	172.7	42.2
12th HSR	2,185	50.8	28.7

years later, in the winter of 1952, all the high school seniors in the state, including many of the 9th grade students tested in 1949, were tested on the 1947 College Edition of the ACE Psychological Examination and Cooperative English Test, Form S, Lower Level, Mechanics of Expression and Effectiveness of Expression. High school percentile ranks (HSR) were procured from the high schools for these seniors. The HSR was based on the senior's scholastic rank in his class at the end of three and one-half years of work.

A sample of 2,185 men and women who had been tested as freshmen was pulled from the files. Correlations were computed between 9th grade total ACE raw score, 9th grade

Table 2

Coefficients of Correlation between 9th Grade Test Scores and 12th Grade Test Scores and High School Percentile Rank \*

Tests	ACE (12th Grade)	Coop. Eng. (12th Grade)	HSR (12th Grade)
ACE (9th Grade)	.80(2169)	.71(2171)	.63(2173)
English (9th Grade)	.75( 681)	.82( 683)	.71( 690)
ACE (12th Grade)		.74(2185)	.65(2185)
English (12th Grade)			.74(2185)

\* In parentheses following the coefficient is given the number of cases upon which each coefficient is based.

total Cooperative English raw score and 12th grade total ACE raw score, 12th grade Cooperative English total raw score and HSR. Table 1 presents the means and standard deviations for each of the variables.

As Table 2 shows, there was a substantial relationship between the 9th grade tests and the corresponding tests given in the 12th grade and with HSR. High School ACE

taken in the 9th grade correlated .80 with College ACE taken in the 12th grade and .63 with HSR. These results indicate the extent to which the high school counselor can interpret 9th grade test scores as predicting high school achievement and 12th grade test scores and can use these predictions to counsel 9th grade students.

*Received February 27, 1953.*



## Selecting Items for Interest Inventory Keys<sup>1</sup>

Kenneth E. Clark and Helen H. Gee

*University of Minnesota*

The use of vocational interest measures assumes that workers in a given occupation have in common certain likes and dislikes, and that these preferences are different from those of workers in other occupations. The extent to which an individual's interest patterns match those of a group is determined by use of a scoring key on an interest inventory. This key is developed by using those responses which are made more frequently by the specific occupational group than by men-in-general (scoring these responses "plus") and those responses made less frequently by the specific occupational group (scoring these responses "minus"). How great the difference in response must be in order for a response to be scored is a difficult question to answer. The difference must be large enough to reduce to a negligible amount the number of chance differences. Yet the number of responses scored must not be so small as to yield a key which is too unreliable for use with individuals. Between these two limits it is still possible to develop many different keys possessing rather widely varying characteristics.

This paper summarizes work which has been done in trying out various methods for the development of scoring keys for the U. S. Navy Vocational Interest Inventory. The reader will note that this work is strictly empirical, although the ideas which are tried out arise from theoretical work. To the extent that interest inventory responses are unique in their psychometric characteristics, the findings of this report are limited in application. It seems reasonable to assume, however, that similar methods of key development would produce similar results when applied to such related measures as per-

sonality inventories, biographical records, and the like.

### Samples Used

Two occupational groups have been used, one civilian and one military. The civilian group is composed of 189 electricians obtained through labor union sources in St. Paul, Minnesota, and, for cross-validation use, 174 electricians similarly obtained in Minneapolis. Keys were developed by comparing their responses with those of members of other occupational groups from St. Paul and Minneapolis. These were: milk wagon drivers, painters, plasterers, bakers, sheet metal workers, printers, warehousemen, plumbers, machinists, shipping clerks, pressmen.

The Navy group is composed of a sample of 261 Aviation Machinist's Mates (AD's) obtained through Receiving Stations on the east and west coasts, and a sample of 292 AD's for cross-validation purposes obtained from the Naval Air Technical Training Command at Memphis. The Navy men-in-general sample used to determine the amount of overlap obtained for various keys is a sample of 200 men drawn randomly from a sample of 1,000 Navy rated men who had been drawn from the total Receiving Station sample in such a way as to reflect the distribution of rates in the Navy as a whole. The entire sample of 1,000 was used to obtain the percentages of responses of men-in-general needed in the development of keys.

### Criteria of a "Good" Key

For purposes of this study, a scoring key is considered good if it does a good job of separating workers in a given occupation from workers-in-general. Thus, a key for Gunner's Mates would perform its function well if the distribution of scores of GM's was markedly different from a distribution of scores of men in another rate, or of men in a variety of different rates. In the following pages, the index of separation of such distributions

<sup>1</sup> The research reported herein was carried out under Contract N6ori-212, T.O. III, NR 151-248, between the Office of Naval Research and the University of Minnesota, and, in part, under a grant from the Graduate School of the University of Minnesota. Able assistance in major parts of the work reported here was given by Mrs. Carolyn C. White and Mr. Norris Ellertson of the project staff.

which shall be used is "percentage overlap." This index gives the number of persons per hundred in one distribution whose scores can be matched by scores in the other distribution. Perfect separation occurs when the highest score in one distribution is lower than the lowest score in the other; in this instance the percentage overlap is zero. No separation at all can be made if the two distributions are identical. When this occurs the percentage of overlap is 100.<sup>2</sup>

A second criterion used in the evaluation of a scoring key is its reliability. In this report reliability is reported as test-retest reliability, obtained by scoring the interest inventories of 90 men students at Dunwoody Industrial Institute, Minneapolis, who took the inventory twice, with an interval of about one month between administrations.

A different sort of criterion which may be used to evaluate methods of scoring the interest inventory would be the relative success of various keys for the prediction of school success or of advancement in rate, or the prediction of re-enlistment, or the prediction of military failure as evidenced by records of disciplinary action or less than honorable discharge. These methods of evaluation are obviously more pertinent for the application of interest inventory scores, but require the passage of a considerable period of time after administration of the inventory, to permit the individual to have a chance to achieve or fail to achieve. Accordingly, these criteria are not used in this report. One might expect that keys which do a good job of separating groups would prove to be the same sorts of keys that would prove useful for these other purposes, but there is, as yet, insufficient evidence to warrant this expectation in the military service. Data have been collected, however, which will give evidence on this point after a sufficient interval of time has elapsed.

In any development of scoring keys based upon empirical methods, there is always the possibility that differences between groups

used to select responses for scoring are chance differences which, upon cross-validation, will tend to disappear. Accordingly, for each of the keys developed and reported upon in this report, a cross-validation sample has been used to determine the amount of regression to be expected. In addition, differences generally have been required for scoring which are large enough to be well beyond the limits within which chance factors would be expected to operate; this method of operating seemed desirable since each key is made up of only a small number of items selected from a total pool of 1140 item responses.

#### Optimal Number of Items in a Scoring Key

Finding no adequate rationale for determining *a priori* the number of item responses to score in developing an occupational key for the vocational interest inventory, attempts were made to make this determination empirically. This work was started with the hope that scoring could be done with less effort than is required with the Strong *Vocational Interest Blank*, which does a good job of separating out occupational groups, but at the expense of a weighting of many item responses to get a score. (Strong assigns weights varying from plus four to minus four to as many as five or six hundred of the twelve hundred possible responses to his blank.)

The first work done to determine how best to develop a scoring key was done with the civilian electrician sample. A series of scoring keys was developed on the basis of the difference in responses of the electrician and other skilled trades groups, as follows: a 6% key was developed by using all item responses with differences in percentage responses of electricians and tradesmen-in-general of six per cent or more. In like manner, a 7% key, an 8% key, a 9% key, and so on, were developed. The series was stopped at a 26% key, when only 21 items remained for scoring.

The comparative merits of each one of these keys may be inferred from the data presented in Table 1. These data indicate the existence of an optimal point in key development, since greatest separation occurs neither at the end of the scale with the smallest number of items, nor at the end of the

<sup>2</sup> This is the index of overlap suggested by Tilton (Tilton, J. W. The measurement of overlapping. *J. educ. Psychology*, 1937, 28, 656-662). Tilton's article provides tables which may be entered using the difference in means for the two distributions divided by the average of their standard errors. For other characteristics of this index, see Tilton's article.



Table 1

Comparison of Various Electrician Scoring Keys in Terms of Overlap and Several Estimates of Reliability

Key	No. of Items in Key	Per Cent Overlap		Test-Retest Reliability
		Original N=189	Cross-Validation N=174	
6%	580	51%	50%	.84
7%	493	49	52	.83
8%	402	47	49	.81
9%	345	48	48	.80
10%	289	44	47	.81
11%	234	46	44	.80
12%	201	46	44	.80
13%	171	44	42	.79
14%	140	44	37	.78
15%	116	41	39	.78
16%	103	41	39	.78
17%	87	39	36	.77
18%	72	40	35	.79
19%	62	38	31	.79
20%	55	37	29	.80
21%	44	37	27	.80
22%	43	37	26	.81
23%	40	39	31	.81
24%	30	42	31	.78
25%	24	48	34	.77
26%	21	50	34	.78

scale with the largest number of items. Keys with smaller numbers of items, in general, are to be preferred. It seems safe to conclude that, as one starts with a small number of items, the addition of more items increases the differentiating power of the key only so long as these items contribute more uniqueness than error; as error increases, the standard deviations of both the criterion and men-in-general groups increase enough to offset the additional increase in mean difference contributed by these items.

With a small number of items, however, some attention needs to be given to problems of reliability. When the only estimates of reliability that were available were made by other than test-retest means, this problem seemed serious enough to warrant the sacrifice of considerable validity in order to achieve minimum reliability. As Table 1 indicates, however, very little is lost in the way of test-retest reliability, by a radical reduction in the number of items scored.

A check on the degree to which this generalization about number of items in the key affecting the validity of the key has been made as part of another study using the Strong *Vocational Interest Blank*. In that study the best key was the one with the smallest number of items scored (24 items). However, these were responses of psychologists, with no sample of answer sheets for men-in-general, so that a different measure of goodness-of-key than percentage overlap was used. No evidence on test-retest reliability of this set of keys was obtained. Even so, it would seem that unit weighting of a fairly small number of items is warranted for scoring of vocational inventory responses.

### Effects of Weighting

While the data on which the decision to use unit weights was based are fragmentary, they indicate clearly that superior separation of groups can be attained by use of such unit weights. Thus, per cent overlap between electricians and tradesmen-in-general was 37% with the best unit-weights key, and was 53% with a key weighted according to the formula used by Strong. The same figures for printers were 40% and 57%, respectively. Scoring the Strong blank, using best unit-weights key, placed men-in-general 3.71 standard deviations below the mean for psychologists in the original sample, and 4.03 standard deviations below the mean for psychologists in the cross-validation sample. Using Strong's method of weighting, men-in-general fell 3.23 standard deviations below the mean for psychologists.<sup>3</sup>

These comparisons do not, of course, indicate that weighting would not improve separation of groups. In fact, the entire literature on multiple regression would suggest otherwise. What they do indicate is that

<sup>3</sup> These data were obtained from sub-samples of responses of psychologists to the Strong *Vocational Interest Blank* used by Kriedt in developing the 1948 Psychologists key. See: P. H. Kriedt, Vocational interests of psychologists, *J. appl. Psychol.*, 1949, 33, 482-488. Kriedt reports that, for the total sample of 1048 psychologists, the means of professional men-in-general and of psychologists are 3.25 standard deviations apart, using the standard deviation of the psychologist group as the unit of measurement (p. 484). Using identical computational methods, the sub-sample above gives a value of 3.23, giving good indication of its representativeness.

a simpler scoring system can separate groups as well as does the more involved method used with the *Strong Vocational Interest Blank*. In the interest of economy of scoring, it thus seems profitable to use unit weights until such time as a real superiority of multiple weights is demonstrated.

#### Heterogeneity of Content of Keys

The selection of item responses for scoring solely on the basis of the percentage difference in response of a reference group and a criterion group will tend, presumably, to give an over-representation of items reflecting certain aspects of the interests of the criterion group, and under-representation of other aspects. Thus, in developing a key for electricians, it might well be that 30 responses indicating a man liked to splice wires, repair circuits, and the like, might be scored, whereas only one response indicating that a man wanted to study in the area of mathematics, electrical engineering, and physics might be scored. Yet both of these kinds of responses are characteristic of the responses of electricians.

In a sense the use of weights might be considered as an attack on this problem. Most weighting is done, however, on the basis of the magnitude of the difference between men-in-general and the specific group, rather than on the basis of the amount of the factor already measured by other item responses. To devise an economical procedure for computing such weights directly would be a genuine contribution. This project has not done so. In the absence of such a procedure, approximation methods must be employed.

The first method employed in this study to improve the composition of a scoring key was an attempt to avoid including in a key too large a number of items reflecting the central core of interests of an occupational group. An iterative method of item selection was therefore employed. First, the best ten items were selected; these were the items on which the responses of the criterion group differed most from the responses of the reference group. All members of the criterion group were then scored for their responses to these items. Another ten items were then selected; for each of these the difference in responses

between reference and criterion groups was still large, and the correlation with the composite of the first ten items was negligible. Another set of ten valid items (i.e., differentiating between criterion and reference groups) which did not correlate with these first twenty was then selected. Finally, ten more valid items unrelated to the first thirty were selected. This key is therefore a fairly heterogeneous key which omits a rather large number of items even though they differentiate members of the occupational group from tradesmen-in-general.

The first groups on which this type of key was tried were civilian electricians. The electrician key which had been developed by simpler means, taking all item responses with a given percentage difference for the criterion and reference group, was already a satisfactory key. The percentage overlap of distributions of scores of electricians and tradesmen-in-general was only 35% in the original group, and 41% in the cross-validation group.

Even so, the use of the iterative method for selecting items for scoring in a key reduced overlap to 30% in the original sample, and to 35% in the cross-validation sample. And this is done without any real drop in the reliability of the key, even though only 40 item responses are scored.

The same comparison of an original key (developed by using all items showing a given minimum difference between criterion and reference groups) and a key developed by iterative methods was made using samples of Aviation Machinist's Mates (AD's) obtained from Navy sources. The AD key developed by original methods is not a very good key in terms of its separation of AD's from Navy men-in-general, since the overlap of these two groups is relatively high—65% for the original group, and 58% for the cross-validation group. Its reliability is, however, rather good. On the other hand, the key developed by iterative methods is a distinctly better key than that developed by original methods when one looks at the overlap between groups, but has a reliability of only .74. These findings are in accord with those obtained with civilian electricians, except that differences are greater between different keys. (The reader should not generalize from these



Table 2

Summary of the Characteristics of Various Scoring Methods Applied to Three Criterion Samples

Group	Type of Key	No. of Items	Per Cent Overlap		Test-Retest Reliability
			Original	Cross-Validation	
Electricians	Original	77	35%	41%	.88
	Iterative	40	30	35	.86
	Gulliksen	63	28	31	.86
Rec. Sta. AD's	Original	83	65%	58%	.85
	Iterative	42	51	51	.74
	Gulliksen	49	56	51	.75

two groups and assume that Navy groups are consistently harder to separate—the AD group was selected because it is a group that gives relatively poor separation from other Navy groups, and hence provides a severe test of the value of the various methods tried with a better-separation group.)

In hopes of developing still better keys at less cost for computation, another type of key was tried. The method of developing this key requires selection of a fairly sizable pool of items, perhaps 100 or more, by taking those items with high validities, and then eliminating those with high indices of internal consistency and only moderate validity. This type of key has been labeled, for want of a better title, the "Gulliksen Key," since the steps taken are similar to those proposed by Gulliksen.<sup>4</sup> Specifically, a key is developed by selecting all items for which the criterion group response differs from that of the reference group by a given amount or more (generally, 12 to 15 percentage points). A large (1,000 for Navy, 550 for civilian groups) men-in-general sample is then scored using this key. The top and bottom 27% of this distribution is used to obtain an estimate of the reliability of each item; the difference in responses of the criterion and reference groups is used as an estimate of the validity of the item.<sup>5</sup> These two values are then plotted against each other much in the manner de-

scribed by Gulliksen (*op. cit.*, p. 384) and items selected much as he recommends. The general effect of the method is to give preference to items which have good validity and which do not correlate highly with other items in the pool.

It should be noted that this method is another approximation method, and is designed to select items having somewhat the same characteristics as the items selected by the iterative method. The Gulliksen method as here used is somewhat easier to employ, is more readily adapted to I.B.M. methods, and hence is more practical than the iterative method. It should also be noted that the values used as estimates of reliability and validity of items differ from those outlined in Gulliksen, since in this analysis gross percentage differences are used in estimating these item characteristics.

The comparison of overlaps and reliabilities of all of these new keys with the original keys developed for electricians and the Navy AD group is summarized in Table 2. In both instances, the Gulliksen key is distinctly superior to the original key in terms of overlap and is perhaps better than the iterative key. The superiority of both methods over the original key is retained in the cross-validation samples as well. In both the electrician and the Navy AD samples this gain seems large enough to warrant the use of the new key in spite of the fact that this key has a lower reliability than the original key.

As noted above, a best unit-weights key for psychologists using the Strong blank resulted in superior separation of psychologists

<sup>4</sup> Gulliksen, H. *Theory of mental tests*. New York: John Wiley & Sons, Inc., 1950. See especially pages 382-385.

<sup>5</sup> Item reliability and validity indices when expressed in correlation terms have yielded, in this work, keys with almost identical characteristics as those obtained using percentage differences.

from men-in-general as compared with a key weighted according to the formula used by Strong. The Gulliksen method described above was also applied to the Strong data but with a slight modification. Since no sample of answer sheets for men-in-general was available it was necessary to base estimates of item reliability on the criterion group. The top and bottom 27% of a subsample of 604 psychologists were accordingly used. A 95-item key resulted from application of the Gulliksen method which differed very little in its effectiveness from the best unit-weights key previously mentioned. Using the Gulliksen key placed men-in-general 3.78 standard deviations below the mean for psychologists as compared with 3.71 for the best unit-weights key. On cross-validation the comparable figures were 3.79 for the Gulliksen and 4.03 for the best unit-weights key. It is to be noted, however, that this best unit-weights key contained only 24 items, and while test-retest reliability is not available, it is doubtful if it would be found to be adequate. This key included all items on which psychologists and men-in-general differed by 33% or more in their responses. The Gulliksen key used items with as low as 18% difference. On a best unit-weights key of 91 items (including all items with 24% or greater difference between psychologists and men-in-general), and in the sense of number of items more nearly comparable to the Gulliksen key, the men-in-general means were 3.26 and 3.42 standard deviations below the means for psychologists on test and cross-validation groups. The implication is clear that item for item, the Gulliksen key results in superior separation, but interpretation must be cautious since information on reliabilities of these keys is not available.

### Summary

The development of a method of scoring responses to an interest inventory so as to maximize the separation of workers in an occupation from workers in general involves consideration of many factors. Taking a cue from applications of multiple regression techniques, we would expect that a point would

be reached when the addition of more items in a scoring key would not be profitable; that, in general, the greater the heterogeneity of item content, the more effective would be the key; and that the use of weighting methods properly applied would increase the degree of separation of groups. Using as criteria of a good key its ability to separate groups (as measured by per cent of overlap of distributions) and its test-retest reliability, it is theoretically possible to demonstrate the importance of each of these points. From a practical standpoint, however, one must determine whether or not approximation methods are usable, and, if so, to what extent these various factors need to be considered when employing these approximation methods.

This report summarizes various methods of developing keys, and provides support for the following statements:

1. When items are scored using unit weights, an optimum number of items can be found for scoring. For the samples used herein, this number seems to be between 40 and 60; when either more or fewer items are scored, the discriminating power of the key is reduced.
2. When item responses are weighted in the manner used by Strong in his *Vocational Interest Blank*, the criterion group is not separated from the reference group as well as when unit scores using the optimum number of items are used. (This is not to say that some weighting system could not be devised which would be superior to unit scoring—obviously such a set of weights could be assigned as to yield a score superior to any score by using multiple regression techniques. What this does say is that the method of weighting used by Strong is not superior to the method of unit weights.)
3. When items are selected so as to increase the heterogeneity of content of a scoring key, the validity of that key is increased, and the test-retest reliability is somewhat decreased. This is true whether items for such a key are selected by an iterative method as described in this report, or by an internal item analysis method.

Received March 23, 1953.



## Practice Effects on the Minnesota Vocational Test for Clerical Workers \*

Howard P. Longstaff

*University of Minnesota*

It is possible that the popularity of a reliable and valid psychological test may become a weakness of that test at least in a given locality. There is some indication that such is the case with the *Minnesota Vocational Test for Clerical Workers*. In the selection of clerical workers this test has been a valuable aid to many business firms in Minneapolis-St. Paul, Minnesota and elsewhere. Its extensive use in the above mentioned Minnesota cities has resulted in many job applicants having taken the test several times. If the test is subject to "practice effects," then the scores made by applicants who have taken it more than once become of questionable value.

Since the *Minnesota Vocational Test for Clerical Workers* has been so widely used, the question has been raised as to what effect practice has upon the scores. Earlier studies indicated a normal practice effect of from 7 to 12 per cent after time intervals of from three to six months (1). This may not seem a prohibitive effect for the time intervals involved, but in actual employment practice much shorter intervals of time are probably the rule. An applicant may apply for a job with several different companies within a matter of hours or days.

The purpose of this study was to measure practice effect on this test over relatively short intervals of time. Two groups of University of Minnesota students in personnel psychology courses served as subjects. Group A was made up of 61 juniors, seniors and graduate students (41 men and 20 women). Group B was comprised of 36 Extension Division students (24 men and 12 women) in an evening class. Group A was given the test successively on a Wednesday, Friday, and Monday, October 1, 3, and 6, 1952.

\* This research was made possible by a grant-in-aid from the Graduate School of the University of Minnesota.

Group B, which met only once a week, was given the test on three successive Monday evenings September 29, October 6 and 13, 1952. The purpose of the study was explained to both groups and they were encouraged to make as much improvement as possible. Since the number of subjects in each group was small and the time intervals between testing were not great, results for the day and night groups are combined.

Table 1 presents the results of the combined groups. It is apparent that considerable practice effects occur. All of the differences between the means are significant at the .1 per cent level. When considered from the standpoint of what these differences mean in terms of centile ranks we observe that the mean scores on the original testing would have had centile ranks on norms for employed clerical workers below 50 while the centile ranks of the mean performance when the test was taken the third time, range from 72 to 91 on the same norms.

A different type of analysis, presented in Table 2, shows much the same thing as the data in Table 1. On trial 3 from 91 to 97 per cent of the subjects reach or exceed the mean score made on trial 1, indicating marked improvement.

As has been shown elsewhere (1, 2, 3, 4, 5) there is a decided sex difference in performance on this test. The subjects in this study behave similarly, as shown in Table 3. Comparing the results of men and women on trial 1 it is apparent that the women are superior to men. It is also obvious that this difference is consistent on successive trials on the test, i.e., comparing trial 1 for men with trial 1 for women, trial 2 for men with trial 2 for women, and trial 3 for men with trial 3 for women. When successive trials for men are compared with the original trial for women the practice effect rather rapidly overcomes the original differences, and by trial 3,

Table 1

Means, Standard Deviations, Differences between Means,  $t$ 's,  $P$ 's,  $r$ 's and Centile Rank the Means Would Have on Norms for Employed Clerical Workers

Part A. Combined Male Groups, N = 65						
Numbers Trials			Names Trials			
1	2	3	1	2	3	
M	118.6	142.5	124.0	154.7	167.3	
S	26.7	25.9	29.3	24.4	23.6	
$\bar{D}_{12} =$	23.9	$\bar{D}_{12} = 33.9$	$\bar{D}_{12} = 30.7$	$\bar{D}_{12} = 43.3$	$\bar{D}_{23} = 12.6$	
$t$	12.0	14.7	18.1	15.5	7.9	
$P$	.001	.001	.001	.001	.001	
$r_{12} =$	.81	$r_{12} = .74$	$r_{12} = .89$	$r_{12} = .67$	$r_{23} = .85$	
Centile rank of mean scores	27	62	47	81	91	
Part B. Combined Female Groups, N = 32						
Numbers Trials			Names Trials			
1	2	3	1	2	3	
M	137.0	157.9	142.7	170.8	178.8	
S	28.4	26.3	30.1	25.7	21.7	
$\bar{D}_{12} =$	20.9	$\bar{D}_{12} = 33.2$	$\bar{D}_{12} = 28.1$	$\bar{D}_{12} = 36.1$	$\bar{D}_{23} = 8.0$	
$t$	8.0	11.5	14.1	10.3	4.0	
$P$	.001	.001	.001	.001	.001	
$r_{12} =$	.86	$r_{12} = .82$	$r_{12} = .93$	$r_{12} = .77$	$r_{23} = .91$	
Centile rank of mean scores	40	70	36	68	81	

77 per cent (Numbers) and 80 per cent (Names) of men scored as high as did the women on trial 1. This is additional evidence of the seriousness of the practice effect on this test.

### Discussion

The practice effect found on the *Minnesota Vocational Test for Clerical Workers* can be explained in part by the nature of the test itself. First, the changed digits in the

Table 2

Percentage of Women and Men (Combined Day and Extension Groups) Who Reach or Exceed the Mean on the First Trial, or the Second Trial or Subsequent Trials \*

Numbers		Names		
Per Cent of Women	Per Cent of Men	Per Cent of Women	Per Cent of Men	
80	85	87	89	Trial 2 vs. trial 1.
94	91	97	97	Trial 3 vs. trial 1.
80	66	69	69	Trial 3 vs. trial 2.

\* Line one of Table 2 shows percentage reaching or exceeding on trial 2 their own mean on trial 1. Line two shows same data for trial 3 compared to trial 1. Line three shows same data for trial 3 compared with trial 2.



Table 3

Percentage of Men (Combined Day and Extension Groups) Who Reach or Exceed the Mean of the Women (Combined Day and Extension Groups) on the Various Trials of Taking the Test \*

Numbers Per Cent of Men	Names Per Cent of Men	
32	30	Trial 1 vs. trial 1.
55	60	Trial 2 vs. trial 1.
77	80	Trial 3 vs. trial 1.
30	30	Trial 2 vs. trial 2.
40	50	Trial 3 vs. trial 2.
30	40	Trial 3 vs. trial 3.

\* Line one of the table shows the percentage of men who on trial one reach or exceed the mean of women on trial one. Line two indicates the percentage of men who on trial two reach or exceed the mean of women on trial one. And line three gives the percentage of men who on trial three reach or exceed the mean of women on trial one. The remaining lines of the table present similar comparisons for trials two with two, trial three with trial two, and trial three with trial three.

number test all occur near the end of the second series of digits. If one catches on to this one can materially improve one's score. Secondly, the items that are changed or are not changed tend to fall into patterns which may help a subject with good visual imagery on repeated trials on the test. Thirdly, on the names part of the test memory may be an important factor in bringing about improvement on successive trials, because the subjects may be able to remember a fairly large number of the name pairs that are changed.

Practice effect is not a new phenomenon; it has been found on other psychological tests besides the *Minnesota Vocational Test for Clerical Workers*. Wherever found, it is likely to be a weakness in any test that is to be used in selecting employees. Especially is this true when it is difficult or nearly impossible to determine how many times a job applicant has taken the test previously. If one

could determine accurately how many times a subject had taken the test and how long a time interval had transpired between testings, correction factors could be worked out. But in the everyday world of employee selection and placement, no reliable method of securing such information exists. Therefore, other ways of overcoming practice effects must be provided if a test subject to practice effect is to have maximum value. The use of alternate forms is one way to reduce this weakness in a test.

### Summary

1. When the *Minnesota Vocational Test for Clerical Workers* is taken successively with short time intervals intervening, marked practice effects occur.

2. With equal amounts of practice the sex difference on test performance remains about constant but with three practice trials men can practically equal the original performance by women.

3. Alternate forms of the test may overcome these weaknesses in the test.

Received March 23, 1953.

### References

1. Andrews, Dorothy M. and Paterson, D. G. *Manual, Minnesota Vocational Test for Clerical Workers*. New York: The Psychological Corporation, 1946. Pp. 5.
2. Loevinger, Jane. *An analysis of verbal and numerical abilities at the junior high school level*. Unpublished Master's thesis, University of Minnesota, 1938.
3. Schneider, Gwendolyn G. Grade and age norms for the Minnesota Vocational Test for Clerical Workers. *Educ. psychol. Meas.*, 1941, 1, 143-156.
4. Schneider, Gwendolyn G. and Paterson, D. G. Sex differences in clerical aptitude. *J. educ. Psychol.*, 1942, 33, 303-309.
5. Thatcher, Meriam. *Sex differences in clerical ability at the fifth and sixth grade levels*. Unpublished paper, Department of Psychology, University of Minnesota, 1940. Pp. 1-12.

## Attitudes Toward Authority-Figures as Correlates of Motivation Among Naval Aviation Cadets<sup>1, 2</sup>

E. P. Hollander and John T. Bair

*U. S. Naval School of Aviation Medicine, Pensacola, Florida*

The interrelationship between attitudes and motivation has already been noted by a number of observers (2, 6, 8, 10). In recent years, evidence of the pertinence of the attitude construct to behavioral criteria has been demonstrated best perhaps in the two-volume series entitled *The American Soldier* (9). Although much of the research reported in these volumes was concerned with service-induced attitudes, large segments of the work dealt with persisting attitudes derived from the serviceman's reference groups external to the service. As a generalization, it might be said that in these studies attitudes were found to be functionally significant in determining the individual soldier's orientation to military life and, accordingly, to his motivation (9, pp. 122-130).

### Problem

This study set forth to determine whether certain attitudes which a Naval Aviation Cadet brings with him to the training program bear a relationship to his level of motivation in training. It is apparent, of course, that attitudes may be ordered in a hierarchy relative to their significance to this particular training situation. That is to say, one would hardly consider that just any attitudes would have significant relevance to motivation in this setting; on the other hand, it is apparent that attitudes toward study or discipline or flying may be of the utmost relevance. In evaluative fashion, then, one might arrive at a grouping of attitudes which are presumed to be of significance in relationship to the motivation of cadets in training. With this

in mind, it was considered that the area of interpersonal attitudes would provide a fruitful area for study. In particular, it was decided that attitudes toward authority-figures, in this case officer-instructors (flight and ground school), would be an appropriate beginning. The intent of the study was to derive implications for further investigations as well as to determine possible applications to selection. The basic hypothesis asserted for test was as follows: that attitudes toward authority-figures would significantly differentiate between cadets of "high" and "low" motivation.

### Procedure

The measurement of attitudes, like the measurement of all psychological variables, offers challenging and oftentimes unique problems. This is especially so where the attitude under scrutiny is both structurally complex and emotionally laden, in this case attitudes toward authority-figures. It soon became apparent that the traditional attitude scale was inadequate and inappropriate to the measurement of an attitude such as this. As a consequence, this technique was discarded in favor of the more flexible open-ended projective questionnaire (7).

The usefulness of this method of attitude-elicitation rests on the fact that it presents the individual with a relatively unstructured stimulus situation in which he may, with equanimity, and without being consciously aware of the process, bring forth feelings that might normally be repressed through social pressures and other forces. Thus, by the employment of this technique, the cadet who felt resentment toward an instructor might vent his feelings without fear of retribution or guilt. The advantage of such a procedure in a military setting is obvious.

In its final form the questionnaire resembled superficially the form developed by Flanagan in his studies of "critical incidents" among Air Force personnel (3). That this was merely a resemblance should be re-emphasized, lest an erroneous impression be conveyed. The main

<sup>1</sup> Opinions or conclusions contained in this report are those of the authors. They are not to be construed as necessarily reflecting the view or the endorsement of the Navy Department.

<sup>2</sup> The authors wish to acknowledge their indebtedness to Dr. Brant Clark for his valuable assistance in the formulation of this report and to Dr. Richard Trumbull, Miss Marjorie Nicholson, and Mr. Calvin Nelson who acted as independent coders of the data.



intent of the investigation was to procure information about instructors only insofar as this information revealed the attitudes of the cadet group under study. The format of the questionnaire was essentially simple. It was presented to the subjects under conditions of anonymity with the inference that only information was being solicited. In addition, subjects were specifically asked not to divulge the names of the individuals about whom they were to write. The cover sheet of this questionnaire contained these instructions: "On each of the following pages you will be asked to write briefly about a person you have known while in the Naval Air Training Program. The instructions indicate that you are to relate just one incident which typified the attitudes and behavior which have led you to make a positive or negative judgment about this person. The incident, however, does not have to be the only one of its kind, nor must it have been the main basis for your evaluation of this person."

On the top of page one the following further instructions were given: "Think of the *best* instructor you had during Pre-Flight or Flight Training. Give *just one* incident which typified the kind of attitudes and behavior which made you feel that he was the best. What were the specific details of his behavior in that particular situation?"

On the top of page two these instructions were given: "Now think of the *worst* instructor you had during Pre-Flight or Flight Training. Here again, give *just one* incident which typified the kind of attitudes and behavior which made you feel that he was the worst. What were the specific details of his behavior in that particular situation?"

Methodologically, two points deserve clarification: first, the "best"-*"worst"* dichotomy was utilized in an effort to secure a degree of polarization of response which would readily yield to differential analysis; second, the instrument was administered under rigorous conditions of anonymity so as to minimize any implied threat.

For purposes of this investigation, motivation was defined operationally. Cadets of "high" motivation were considered to be those who had successfully completed the basic flight stage of the Naval Air Training Program.<sup>3</sup> Cadets of "low" motivation were

those who voluntarily withdrew from the program during this stage.

During a three months period in the fall of 1951, the questionnaire was administered to a total sample of 137 cadets classified as follows: 72 cadets who were leaving training at their own request (the "low" motivation group) and 65 cadets who had successfully completed basic flight training (the "high" motivation group). In both instances administration of the questionnaire was part of a routine check-out procedure and was usually carried on with small groups numbering five or less.

A summary comparison of the two criterion groups will be found in Table 1. With respect to age and active duty time before entering training they were quite comparable. On the whole, however, cadets dropping at their own request tended to have a significantly greater amount of formal education prior to training. This latter finding corroborates, in part, certain of the results growing out of a previous report from this command (1).

Following the administration procedures, responses to the questionnaire form were abstracted so as to yield only core phraseology relevant to the instructor's behavior and the cadet's reaction to this behavior. These abstracts were thereupon transcribed on 3 × 5 cards and assigned code numbers at random so as to eliminate insofar as was possible subjective bias in the content analysis procedure which followed. Thus, at no time during the categorization of this data did the judges know the disposition of the cadet whose response was in hand.

As a next step, all of the responses to the two instructional "sets," that is, "best" and "worst" instructor, were sifted to secure descriptive elements of behavior. From these, a number of categories of behavior were developed, which subsumed behavioral elements of similar quality and as much as possible used the language of the respondents rather than that of the investigators. In every instance, these categories were developed independently of one another in terms of an either-or criterion. That is, either the behavior was described in the response or it

<sup>3</sup> The program is divided into three major phases: Pre-Flight, Basic Flight, and Advanced Flight. In virtually all cases, cadets who have completed Basic have been in training for one year or more. By this time attrition is minimal and the likelihood of success is very high.

Table 1

Summary Comparison of Motivation Criterion Groups with Regard to Age, Previous Education, and Previous Military Service

Group	Age (Years)		Education* (College Semesters)		Previous Active Military Duty (Months)	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
High Motivation (Successful) N = 65	22.0	1.5	3.3	2.6	21.8	13.2
Low Motivation (Withdrawal) N = 72	22.2	1.4	5.5	2.5	20.6	14.0

\* The t test of significance for the difference between the means for the college education variable was found to be 4.89. This is significant at the 1% level of confidence.

was not. Thus, overlap was possible and, indeed, very frequently took place—but only as a result of the respondent's having mentioned more than one major behavioral element.

As a check on reliability of judgment, three independent coders were asked to discriminate one category of behavior, for each of the two instructional sets, within the total population of responses to that set. Percentages of agreement with the principal investigators and the three independent coders were computed for the response categories selected for the reliability check. All were found to reach an acceptable level.<sup>4</sup>

### Results

Frequency of response under each major category for the two cadet groups was subjected to a chi-square analysis. Table 2 presents the findings of this procedure comparing the major categories of "best" and "worst" instructor behavior for the successful and withdrawal cadet responses. In general, Table 2 reveals that the cadets of high motivation tend to manifest attitudes toward the interpersonal quality of instructor behavior while those of low motivation, on the other hand, tend to show attitudes directed at the instructor's success or failure in his role as a teacher. Close scrutiny of this table indicates that under the "best" instructor set,

cadets of the "high" motivation group responded with significantly greater frequency within the categories of *personal interest* and *patience* than did the cadets in the "low" motivation group. On the other hand, the "low" group, under this same set, responded with significantly greater frequency than did the "high" group within the categories *good instructional techniques* and *extra help*. Under the "worst" instructor set, the "high" motivation group reacted with significantly greater frequency than the "low" group within the category *verbal assault* and with significantly less frequency than the "low" group within the category *poor instructional techniques*. No significant difference between the groups was found within the *indifference* category, under this set.

### Discussion

The results indicate that differences of attitude toward authority-figures do exist between cadets of "high" and "low" motivation. The hypothesis, therefore, was substantiated. Specifically, it would appear that there is a degree of variation in identification with instructors between cadets of the two criterion groups. Indeed, it may be that this process of identification may account for the differences obtained.

While it was initially considered that the attitudes studied here were brought by the cadets to the training program, one might properly question the actual temporal relationships involved. That is, were the attitudes toward these authority-figures brought

<sup>4</sup> Two response categories were checked. The percentages of agreement for the response category *patience* were .96, .87, and .88 for each of the independent coders; for *verbal assault* the percentages were .95, .94, and .93, respectively.



Table 2

Chi-Square Analysis and Significance Levels between the Motivation Criterion Groups for the Major Response Categories

"Best" Instructor				
Response Category	Per Cent*		$\chi^2$	P
	High Group (N=65)	Low Group (N=69)		
Showed Personal Interest	75	44	12.91	<.001
Indicated Patience	43	20	8.08	<.01
Used Good Instructional Techniques	37	63	9.65	<.01
Gave Extra Help	9	26	6.49	<.02
"Worst" Instructor				
Response Category	Per Cent*		$\chi^2$	P
	High Group (N=62)	Low Group (N=70)		
Manifested Verbal Assault	61	21	21.74	<.001
Used Poor Instructional Techniques	18	40	7.83	<.01
Indicated Indifference	42	37	1.22	>.30

\* The N's given here for the groups represent the actual number of people in the criterion groups who responded to the "set."

to the training situation, or were they conditioned mainly by experiences in training? A research project has recently been completed (5) which essentially duplicated the current investigation in order to provide an answer to this question. In this study, cadets *just entering training* were given a similar questionnaire form in which they were asked to give parallel information on previously encountered authority-figures, that is, high school or college instructors. The results of this study indicate quite conclusively that attitudes of the cadets who subsequently withdrew from training were similar to those of the "low" motivation group of the present study. This group tended to describe the skill or lack of skill of their high school or college instructor in his role as a teacher at a significantly higher level than did the cadets who remained in training. Thus, it appears that attitudes toward authority-figures are among the attitudes persistently held by the cadets, and are related to their level of motivation in the Naval Air Training Program. A number of related investigations

designed to articulate this relationship still further are now being conducted. On the whole, it would seem that this attitude-elicitation technique bears further scrutiny as a possible device for the assessment of motivation in a number of settings.

### Summary

This paper reports on attitudes toward authority-figures which discriminated between Naval Aviation Cadets of "high" and "low" motivation. The "high" motivation group consisted of 65 cadets who had successfully completed Basic Flight Training, and the "low" group consisted of 72 cadets who were withdrawing from training voluntarily. Both groups were required to complete anonymously an open-ended questionnaire form which required them to describe a sample of behavior characteristic of their "best" and "worst" officer-instructors (ground and flight). Content analyses were undertaken and frequencies for each content category were determined for both groups. The results revealed that cadets of "high" motiva-

tion tended to manifest attitudes concerning *interpersonal relationships* with their officer-instructors while the "low" group stressed *competence of the instructor in his role as a teacher*. Interpretations were suggested with respect to cadet identification with authority-figures as a motivational factor in this setting.

Received April 23, 1953.

### References

1. Bair, J. T. *Non-test predictors of attrition in the Naval Air Training Program*. Project Number NM 001 058.05.02. U. S. Naval School of Aviation Medicine, Pensacola, Florida, 28 April 1952.
2. Cantril, H. *The psychology of social movements*. New York: John Wiley, 1941.
3. Flanagan, J. C. *AAF Aviation Psychology Program, Research Report No. 1*, Washington, D. C., U. S. Government Printing Office, 1948.
4. Hollander, E. P. and Bair, J. T. *The significance of attitudes toward authority-figures in discriminating between Naval Aviation Cadets of "high" and "low" motivation*. Project Number NM 001 058.05.03. U. S. Naval School of Aviation Medicine, Pensacola, Florida, 27 May 1952.
5. Hollander, E. P. and Bair, J. T. *Pre-Training attitudes toward authority-figures as predictors of inadequate motivation among Naval Aviation Cadets*. Project Number NM 001 058.05.05. U. S. Naval School of Aviation Medicine, Pensacola, Florida, 10 November 1952.
6. Newcomb, T. *Social psychology*. New York: The Dryden Press, 1950.
7. Saenger, G. and Proshansky, H. Projective techniques in the service of attitude research. *Personality*, 1950, Symposium No. 2, pp. 23-24.
8. Sherif, M. *An outline of social psychology*. New York: Harper, 1948.
9. Stouffer, S. A., et al. *The American soldier*. (Vols. I and II), Princeton: Princeton University Press, 1949.
10. Thomas, W. I. A theory of social personality, Chapter IX in *Social behavior and personality*. (Edited by E. H. Volkart.) New York: Social Science Research Council, 1951.



## Readability of Employee's Letters in Relation to Occupational Level

Arthur C. MacKinney and James J. Jenkins

*University of Minnesota*

In any form of written communication it is obviously of great importance to have information concerning the reading ability of the audience and to use that information in the communication process. Since 1948 (10), a large number of articles have appeared in the psychological literature and elsewhere stressing the need to simplify and make more readable the communications which managements direct to their employees. (See the bibliography by Hotchkiss and Paterson [8].) These articles have suggested the use of readability formulas (most popularly those presented by Flesch) as one means of controlling the level of the communication and thus attaining the goal of better understanding. Many writers, following the lead of Flesch (5, 6) have used educational achievement as a base from which the comprehension ability of the audience may be estimated. Other writers have contributed the results of reading comprehension tests given to selected samples of special audiences. Here, for example, one finds the work by Bellows and Palmer (1) and Colby and Tiffin (2) on the reading levels of foremen and supervisors. For the most part, however, some indirect estimation procedure must be used since the results of applying reading comprehension tests to rank-and-file employees in industry are not available.

This paper has a two-fold purpose, first, to advance tentatively another estimation procedure and, second, to consider in its own right the data revealed by this technique. It was hypothesized that the readability level of employee-written communications should reflect the effective literacy level of the employees. It was further hypothesized that literacy level increases (as does education and intelligence) with higher occupational levels. This would mean, then, that the readability difficulty of employee-written letters as measured by the Flesch formula should increase

as occupational level increases. Briefly, it was our belief that *in general* the complexity of one's writing provides an indirect index to the complexity of material which one can readily comprehend and that, since it is generally agreed that reading ability increases with occupational level, complexity of writing will increase also.

### Method

A total of 400 employee-written letters were made available from the General Motors "My Job Contest" (Evans and Laseau [3]).<sup>1</sup> These letters were randomly drawn from a 10 per cent sample of the 174,854 letters received in this contest. While these letters are not "typical" writing samples from the employees, they are letters written under standard stimulus conditions and hence are uniquely comparable.

Average sentence length, syllable counts, and Flesch Reading Ease scores were determined for each of the letters on the basis of a 100-word sample from each letter. In 67 instances the letters contained less than 100 words, so the RE scores were determined by prorating these on the basis of the total words available in that letter. The average length of these prorated letters was 71 words. All counting was done independently of salary level information.

It is to be noted in connection with this analysis that the determination of average sentence length for use in the Flesch RE formula is done on the basis of separate ideas, independent of punctuation (which was of dubious accuracy at best in these letters). This admittedly could introduce a source of error since a change of one sentence

<sup>1</sup> The writers wish to express their appreciation for the cooperation of the Employee Research Section, General Motors Corporation and especially to Dr. Chester E. Evans. The 400 employee letters and the occupational descriptions used in this article were furnished by that organization.

in the 100-word sample changes the average sentence length and the RE score markedly. However, the reliability of the Flesch RE measures has been shown to be quite satisfactory (7).

Following the determination of the RE scores, letters were then classified by occupational level of their writers. There were two major groupings, the salaried and the hourly employees.

The salaried group included the "skilled group with responsibilities added," the "skilled group," and the "partially skilled group." Originally the salaried group included "learners" but this category was eliminated because of the small number of cases. The salary group was generally defined as follows:

"Sub-managerial and clerical occupations involving supervising, coordinating, guiding, and performance of general clerical work. Primarily concerned with preparation, transcription, systematizing and filing of oral and written communications in offices, shops, and other places where such functions are performed."

The group of hourly employees was divided in accordance with the traditional classification into "skilled," "semiskilled," and "unskilled." These were defined as follows:

"*Skilled*: Includes craft and manual occupations that require predominantly a thorough and comprehensive knowledge of processes involved in the work, exercise of considerable independent judgment, usually a high degree of manual dexterity, and, in some instances, extensive responsibility for products and equipment. Employees in these occupations often become qualified through apprenticeship or extensive training periods."

"*Semiskilled*: The exercise of manipulative ability of a high order within a fairly well-defined work sequence. The major reliance, not so much upon the employee's judgment or dexterity, but vigilance and alertness, in situations in which lapses in performance would damage equipment or product. These occupations may require the limited performance of part of a craft or skilled occupation."

"*Unskilled*: Manual occupation involving performance of simple duties which can be learned in a short period of time. Little or no independent judgment is required and such occupations require no similar job experience."

Some letters were dropped from the sample at each stage of the analysis. In all, 26 cases were discarded leaving a total of 374 letters for final analysis. As stated before, letters by "learners" were discarded. Occupational classifications were not available or were in doubt for several of the letters.

### Results

Mean and standard deviation of RE scores were calculated for each of the occupational groups. These are presented in Table 1. Analysis of variance applied to the means of these groups yielded an F value of 10.61 which is, of course, significant far beyond the .01 level.

An inspection of Table 1 reveals that a clear hierarchy of Mean RE scores is not only evident between the major groups but within them as well. The means for the "skilled" salaried people places them at Flesch's "Fairly Difficult" level, typical of a quality magazine, indicating reading achievement levels from 10th to 12th grade and requiring some high school for understanding. The "partially skilled" salaried employees and the "skilled" hourly employees write at a mean level equivalent to the digests, "Standard," indicating reading achievement within the 8th and 9th grade levels which requires the completion of 7th or 8th grade for un-

Table 1  
Reading Ease Scores of Employee Letters

Occupational Classification	N	Mean	S.D.
Salaried Employees:			
Skilled with Responsibilities	18	53.7	10.8
Skilled	17	53.6	15.5
Partially Skilled	21	61.7	14.0
Hourly Employees:			
Skilled	51	64.0	12.9
Semiskilled	218	69.1	14.1
Unskilled	49	72.9	12.1



Table 2  
Percentage of Employees in Each Occupational Group Writing at Each Reading Ease Level

Occupational Classification	Flesch Reading Ease Levels								Total
	N	VD 0-29	D 30-49	FD 50-59	S 60-69	FE 70-79	E 80-89	VE 90-100	
Salaried:									
Skilled with Responsibilities	18	...	44.4	27.8	22.2	5.6	...	...	100.0
Skilled	17	...	41.2	17.6	17.6	23.5	...	...	100.0
Partially Skilled	21	...	19.0	23.8	23.8	28.6	4.8	...	100.0
All Salaried	56	...	33.9	23.2	21.4	19.6	1.8	...	100.0
Hourly:									
Skilled	51	2.0	11.8	25.5	25.5	25.5	5.9	3.9	100.0
Semiskilled	218	0.5	11.9	13.3	18.3	30.7	20.2	5.0	100.0
Unskilled	49	...	2.0	14.3	18.4	36.7	20.4	8.2	100.0
All Hourly	318	0.6	10.4	15.4	19.5	30.8	17.9	5.3	100.0
All Employees	374	0.5	13.9	16.6	19.8	29.1	15.5	4.5	100.0

understanding according to Flesch. The "semi-skilled" and "unskilled" hourly workers write at a mean level which is like slick fiction, "Fairly Easy," indicating a reading achievement of 7th grade and requiring completion of 6th grade for understanding.

More revealing is the tabulation of the percentages of persons of each occupational level who wrote at each of Flesch's readability levels. These data are summarized in Table 2.

Here again the progression of reading ease scores over the occupational level hierarchy is striking. For example, it may be seen that 44 per cent of the "skilled with responsibilities"

ties" salaried group write at the "Difficult" level while only 2 per cent of the unskilled (hourly) group write at this same level.

To further facilitate use of these data, the percentages of Table 2 were cumulated for each occupational level. The results are presented in Table 3.

### Discussion

The data, as presented, are of descriptive interest just as they stand. However, a crucial question remains. Since these letters are samples of employees' *writing*, does this really indicate the reading comprehension level of

Table 3  
Cumulative Percentage of Employees in Each Occupational Group Writing at Each Reading Ease Level

Occupational Classification	Flesch Reading Ease Levels							
	N	VD 0-29	D 30-49	FD 50-59	S 60-69	FE 70-79	E 80-89	VE 90-100
Salaried:								
Skilled with Responsibilities	18	...	44.4	72.2	94.4	100	100	100
Skilled	17	...	41.2	58.8	76.4	100	100	100
Partially Skilled	21	...	19.0	42.8	66.6	95.2	100	100
All Salaried	56	...	33.9	57.1	78.5	98.1	100	100
Hourly:								
Skilled	51	2.0	13.8	39.3	64.8	90.3	96.2	100
Semiskilled	218	0.5	12.4	25.7	44.0	74.7	94.9	100
Unskilled	49	...	2.0	16.3	34.7	71.4	91.8	100
All Hourly	318	0.6	11.0	26.4	45.9	76.7	94.6	100
Total	374	0.5	14.4	31.0	50.8	79.9	95.4	100

these same employees. There is no rigorous answer to this question at the present time. A consideration of the writing process as opposed to the reading process, production of a word as opposed to its recognition, the special conditions of a contest with very substantial prizes, the special pressures on the individual to make some kind of an entry so his group might receive a participation award, and the possibility that many of the letters were written with help from members of one's family, neighbors, etc., preclude any realistic discussion of whether an individual writes at a level higher than, lower than, or similar to his reading comprehension level.

Some supporting evidences, however, incline the writers to the view that this is representative writing and that it is indicative of minimal reading skill. First, repeated analyses of house organs (presumably written by salaried employees who are "skilled with responsibilities") show their mean level to be very close to that indicated in this study. In general, their writing averages RE scores of about 50 (4, 11) as compared to the average of 54 obtained in this study. Second, the study by Bellows and Palmer (1) of reading comprehension of foremen (who are presumably like the "skilled" hourly worker) seem to match very closely the data obtained in this study for this group. Their data are presented in modified form in Table 4 for comparison with the group from this study. Colby and Tiffin (2) find the median reading grade for factory supervisors to be the 10th grade level while this study shows a median in the 9th grade level.

If one accepts the data from these letters as reflecting the minimal effective literacy level of the employees, then this industrial audience has been somewhat more clearly delineated. To reach 95 per cent of all employees for example, Table 3 indicates it would be necessary to write at the "Easy" level of 80 to 90 (pulp fiction). This is the level which Flesch has predicted would reach 91 per cent of the adult population.

On the other hand, if one were concerned only with reaching the top salary-level group represented here (skilled with responsibilities added) 94 per cent of that group would find

Table 4

Reading Comprehension Grade of Foremen as Measured by Bellows and Palmer in Comparison with Estimated Comprehension Level of Skilled Sample in this Study

Reading Comprehension Grade Level	Per Cent of Foremen (Bellows and Palmer) N = 100	Per Cent of Skilled (Estimated from RE Scores) N = 51
16+	4	2.0
13-16	21	11.8
10-12	26	25.5
8-9	27	25.5
7	6	25.5
6	8	5.9
4-5	8	3.9

the "Standard" level within their reading comprehension. This would be a Reading Ease level of 60 to 70 and is typical of digest magazines.

Writing at the "Fairly Easy" level (typical of slick fiction; Reading Ease 70 to 80) would be easily understood by 80 per cent of all employees. It would be well within the grasp of almost all salaried employees. However, only 71 per cent of the unskilled employees would readily comprehend this "Fairly Easy" level of writing.

It is interesting that this standard of RE of 70 or easier was recommended by Paterson and Walker (11), Farr, Paterson, and Stone (4), and by Lauer and Paterson (9) in their studies of industrial communications intended for "rank-and-file employees."

### Summary

A total of 400 employee letters were randomly drawn from the 10 per cent sample of letters received in the General Motors "My Job Contest." One 100-word sample from each letter was analyzed by the Flesch Reading Ease formula. The letters were then sorted by occupational level of the writer. The mean RE score and the standard deviation were computed for each of six occupational levels. Mean differences between the groups were highly significant.

A hierarchy of mean RE scores was found to exist ranging from a mean of 54 (Fairly



Difficult) for the "skilled" salary groups to a mean of 73 (Fairly Easy) for the "unskilled" hourly employees. A table showing the percentage of each group writing at each RE level was prepared to more fully describe the distributions. Some evidence suggesting that the writing was representative and indicative of comprehension level was presented. The results were interpreted as confirming previous readability studies of industrial communications and as providing a guide for the preparation of industrial communications.

Received April 2, 1953.

### References

1. Bellows, R. M. and Palmer, D. H. Unpublished study. Cited in Bellows, R. M., *Psychology of personnel in business and industry*. New York: Prentice-Hall, 1949, p. 499.
2. Colby, A. N. and Tiffin, J. Reading ability of industrial supervisors. *Personnel*, 1950, 27, 156-159.
3. Evans, C. E. and Laseau, L. N. *My job contest*. *Personnel Psychol. Monogr.*, 1950, No. 1.
4. Farr, J. N., Paterson, D. G., and Stone, C. H. Readability and human interest of management and union publications. *Industr. Labor Relat. Rev.*, 1950, 4, 88-93.
5. Flesch, R. F. *The art of plain talk*. New York: Harpers, 1946, p. 210.
6. Flesch, R. F. *The art of readable writing*. New York: Harper, 1949, p. 499.
7. Hayes, Patricia M., Jenkins, J. J., and Walker, B. J. Reliability of the Flesch readability formulas. *J. appl. Psychol.*, 1950, 34, 22-26.
8. Hotchkiss, S. N. and Paterson, D. G. Flesch readability reading list. *Personnel Psychol.*, 1950, 3, 327-344.
9. Lauer, Jeanne and Paterson D. G. Readability of union contracts. *Personnel*, 1951, 28, 3-7.
10. Paterson, D. G. and Jenkins, J. J. Communication between management and workers. *J. appl. Psychol.*, 1948, 32, 71-80.
11. Paterson, D. G. and Walker, B. J. Readability and human interest of house organs. *Personnel*, 1949, 25, 438-441.

## Scaling Procedures Based on the Method of Paired Comparisons

Sam L. Witryol

*Department of Psychology, The University of Connecticut*

The primary purpose of this paper is to present an experimental comparison of three scaling approaches to the method of paired comparisons: Thurstone's Case III and Case V, and Guilford's Short Cut. Recent literature pertaining to a variety of related practical and theoretical developments will also be briefly reviewed.

There appears to be a resurgence of interest on the part of investigators in many areas of psychology in the applicability of the method of paired comparisons to practical scaling problems. Fortunately some important original contributions, clarifying measurement problems and re-examining basic assumptions, have also recently been published. The work of Mosteller (21, 22, 23, 24) is exemplary and constitutes, in the opinion of the writer, the most brilliant rational discussion of paired-comparison scaling features since Thurstone's early developments (26, 27, 28).

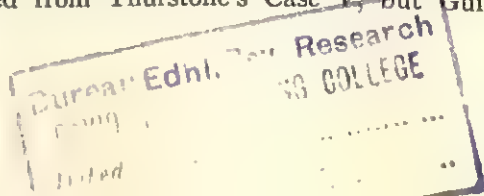
In a previous investigation (30), the writer employed Thurstone's Case V for scaling paired-comparison data on teacher-generated motivational values in the classroom. The Case V scale values from four different classroom groups were compared with the values from the same samples scaled by Case III. The correlations between these scale values obtained by these two methods from four sets of data were essentially unity, despite the fact that one of the assumptions for Case V (equal discriminial dispersions) appeared to have been violated. In the present investigation the values obtained from these data by means of Thurstone's Case III and Case V approaches were compared with those obtained by Guilford's Short-Cut method (10). The Guilford method was rejected by the writer in the earlier study because it seemed to lack a defensible rationale for the discriminial unit. This decision will be re-examined here in the light of experimental findings from the present study and from related researches.

Gulliksen (11) has discussed the broad scaling characteristics and power of the method of paired comparisons, and Burros (4) has pointed out that this "psychophysical" procedure has special value for scaling stimuli when the "physical" correlates are not easily discernible. Coombs (5) has noted that the data in most scaling experiments are qualitative in nature. With these considerations in mind, it is worthwhile to evaluate the method of paired comparisons in terms of generality of measurement to various types of qualitative data and also in terms of economy of application.

### Previous Research Findings

In the present writer's earlier investigation (30), the scaling rationale developed by Thurstone for Case III and for Case V, and the rationale developed by Guilford for his Short-Cut method were described in some detail. The major relevant features will be briefly reviewed here. Thurstone (26) examined various assumptions under five special cases for the application of the method of paired comparisons to scaling the psychophysical law of comparative judgment. The choice of these Thurstone scaling methods generally hinges upon the selection of either Case III or Case V, based upon whether the stimulus dispersions are approximately equal or unequal. Thurstone admitted that these dispersions could never be directly observed (27), and later developed a statistical procedure for approximating the measurement of the "ambiguity of each stimulus" (28). The important practical consideration was the fact that Case V, a much simpler scaling device, was indicated by Thurstone to be applicable if the stimulus dispersions were approximately equal.

As a laborsaving alternative to Case V, a Short-Cut method was devised by Guilford (8). The results with this procedure yielded very high correlations with the results obtained from Thurstone's Case V, but Guil-





ford was unable immediately to present an adequate mathematical and psychological justification. Later, he attempted this latter task, but he could not develop a unit for the psychological scale (9).

Recent empirical findings have been suggestive. Edwards (6) reported a very high correlation, approaching unity, between values obtained with Case III and Case V, but did not indicate the values for the stimulus dispersions. Koch<sup>1</sup> (17) also reported very high correlations (about +.99) between scale values obtained with Case V and Guilford's Short Cut. Satter (25) found Guilford's approach to be a highly reliable method for job evaluation. These findings are in general consistent with the research experience (30, 31, 32) of the writer as well as with the experimental findings to be reported in the present investigation.

However, an adequate rationale is not readily ascertainable from a review of the empirical findings, and one must turn to the recent brilliant efforts of Mosteller (21, 22, 23, 24) for the most productive and provocative leads. Mosteller presents a careful mathematical rationale which makes it possible to relax restrictions previously considered basic assumptions for the application of Case V. Thus he has demonstrated that:

1. An assumption of equal correlations, as well as one of zero correlations, between the stimulus pairs is tenable (21).
2. An aberrant stimulus standard deviation affects only the position of that stimulus involved (22).
3. If the aberrant stimulus dispersion is near the center of the scale, scale positions of the other stimuli will not be seriously affected (22).
4. The requirement of normality in the original distribution is not necessary (24).

Furthermore, Mosteller (23) proposed a test of goodness of fit of observed to theoretical proportions; this method is also designed to test unidimensionality.

A recent rational development by Burros (4) is noteworthy. He worked out a method

for estimating stimulus dispersions. His results compare favorably with Thurstone's as valid estimates, and they have the advantage of requiring less arithmetical computation, although Burros' formulae are more complicated.

The problem of unidimensionality of the paired-comparison scale has received serious consideration. Most investigators have applied Thurstone's methods to data assumed to be ordered along a single dimension. However, Gulliksen's excellent analysis (11) demonstrated the feasibility of the application of the method of paired comparisons to multidimensional scales. In fact, he reasoned that this power of the paired-comparison method was a significant advantage over ordinary ordinal scales, and he reviewed researches which were exemplary of these possibilities. In any event, determination of unidimensionality or multidimensionality is an important factor in a specific experimental situation.

Mosteller (23), as noted above, has developed a chi square test for unidimensionality with the restrictive assumption relaxed to equal in addition to zero correlations between the stimulus pairs. Kendall and Smith (14, 15) derived a "coefficient of agreement" (also "coefficient of consistence") to test the assumption of linearity of the paired-comparison variate under consideration. Johnson (13) described this test, and, recently, Balinsky, Blum, and Dutka (3) demonstrated its applicability in determining the consistency of product preferences. Finally, an experimentally provocative and potentially fruitful approach to multidimensional variates was suggested by Andrews (1). He performed a factor analysis of the multidimensional elements in stimuli presented in paired-comparison form; his analysis was derived from the table of proportions conventionally calculated as part of the computational process.

#### Experimental Procedure

The paired-comparison data analyzed in this experiment were obtained in an earlier investigation (30) where the methodological details were fully described; the main features will be briefly reviewed. The stimuli consisted of a group of ten praiseworthy and

<sup>1</sup> Obtained in part by personal communication from Dr. Helen L. Koch, University of Chicago, Oct. 4, 1950.

of another group of ten blameworthy categories derived from teacher-generated motivational values as reported by school children. Each of these two groups of ten stimuli were presented in paired-comparison form to 1,120 school children in grades 6-12. The subject's task was to judge which of each pair of stimuli was more teacher-approved or disapproved. Case V scale values were calculated from the responses to these stimuli for each sex by age-grade classification, so that each sample population included 80 subjects. Thus, a total of 28 sets of scale values, with ten stimuli in each, were computed.

For purposes of the present experiment four sample sets from the above data were selected for comparative analyses by means of three different scaling procedures: Thurstone's Case V and Case III, and Guilford's Short Cut. The sets were selected from the total population in such a manner as to represent both sexes, both experimental conditions (praise and blame) and, finally, different age-grade levels. Each sample represented a particular sex, experimental condition, and age-grade level. The specific nature of the sample sets can be readily observed from the captions of the tables and figures in the results, below.

### Results

The scale values obtained by each of the three scaling approaches to paired comparisons are presented in Tables 1, 2, 3, and 4. The discriminial dispersions of each of the stimuli, as estimated by Thurstone's Case III, are shown in the last column of each table. Twelve product-moment correlations obtained by comparing the scaling results calculated by the three different approaches range from .987 to .999; these intercorrelations appear in the bottom three rows of the four tables. The averages of the four intercorrelations obtained by comparing Case V with the Short Cut, Case III with Case V, and Case III with the Short Cut in all the samples are .998, .994, and .991, respectively.

It should be noted from the tables that the discriminial dispersions in the last columns are not approximately equal. It can be seen by inspection that there is a considerable range in these estimated dispersions in each

Table 1  
"Teacher Praise" Scale Values Computed by  
Thurstone's Case III and Case V and  
Guilford's Short-Cut Methods  
(80 Sixth-Grade Boys)

Behavior-Activities (Stimulus)	Case III	Case V	Short Cut	Discriminal Dispersions (Estimated by Case III)
Honest	1.21	1.08	.82	1.414
Obey	.73	.66	.49	.850
Attention	.70	.63	.48	.911
Polite	.68	.60	.47	.660
Cooperative	.38	.35	.28	1.399
Talking	.19	.21	.18	.848
Industrious	.14	.15	.16	.827
Help	.13	.14	.13	.955
Independent	.04	.04	.05	.939
Clean	0	0	0	1.195
$\sigma$	.375	.329	.242	
$r_{III-V}$	.999			
$r_{III-so}$	.997			
$r_{V-so}$			.999	

of the four samples. Finally, the standard deviations of the scale values obtained by each of the three approaches are systematically smaller in Case V and in the Short Cut, respectively, than in Case III.

Table 2  
"Teacher Praise" Scale Values Computed by  
Thurstone's Case III and Case V and  
Guilford's Short-Cut Methods  
(80 Twelfth-Grade Girls)

Behavior-Activities (Stimulus)	Case III	Case V	Short Cut	Discriminal Dispersions (Estimated by Case III)
Honest	3.06	2.96	1.82	1.318
Polite	2.58	2.51	1.44	.786
Industrious	2.17	2.10	1.20	1.127
Attention	2.14	2.06	1.15	.595
Independent	1.93	1.92	1.12	1.019
Cooperative	1.90	1.90	1.11	1.351
Obey	1.81	1.82	1.03	.728
Talking	1.26	1.11	.60	.823
Clean	1.20	1.06	.54	1.274
Help	0	0	0	.980
$\sigma$	.798	.791	.482	
$r_{III-V}$	.998			
$r_{III-so}$	.992			
$r_{V-so}$			.997	



Table 3

"Teacher Scold" Scale Values Computed by  
Thurstone's Case III and Case V and  
Guilford's Short-Cut Methods  
(80 Eighth-Grade Girls)

Behavior- Activities (Stimulus)	Case III	Case V	Short Cut	Discriminal Dispersions (Estimated by Case III)
Rude	1.88	1.67	1.06	.685
Dishonest	1.79	1.58	1.02	1.346
Disobey	1.73	1.48	.93	.925
Disturb	1.56	1.26	.772	.551
Chew Gum	1.49	1.19	.770	1.281
Fight	1.34	1.07	.69	1.184
Poor Work	1.24	.96	.63	.589
Attention	1.14	.83	.54	.910
Talking	1.09	.80	.50	1.781
Untidy	0	0	0	.747
$\sigma$	.512	.461	.292	
$r_{III-V}$	.989			
$r_{III-80}$	.988			
$r_{V-80}$			.999	

These quantitative results are graphically represented in Figures 1, 2, 3, and 4.

### Discussion

The empirical comparisons in this experiment suggest the following conclusions:

1. The Case V approach appears to yield essentially the same scale distribution as the

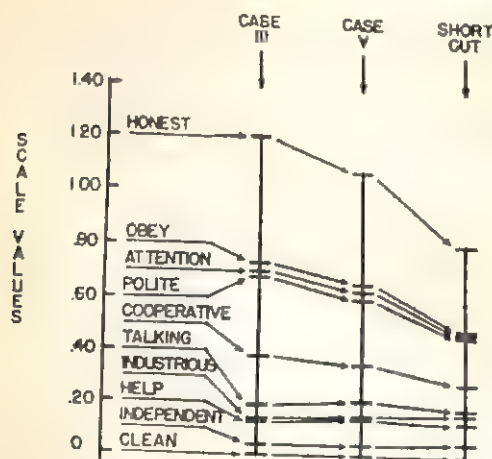


FIG. 1. "Teacher praise" scale values computed by Thurstone's Case III and Case V and Guilford's Short-Cut methods (80 sixth-grade boys).

Table 4

"Teacher Scold" Scale Values Computed by  
Thurstone's Case III and Case V and  
Guilford's Short-Cut Methods  
(80 Tenth-Grade Boys)

Behavior- Activities (Stimulus)	Case III	Case V	Short Cut	Discriminal Dispersions (Estimated by Case III)
Dishonest	1.95	1.76	1.18	1.352
Rude	1.51	1.42	.96	1.159
Disobey	1.49	1.40	.92	.757
Fight	1.29	1.16	.74	.939
Chew Gum	1.25	1.12	.70	1.124
Disturb	1.23	1.10	.67	.855
Poor Work	1.08	.90	.59	.605
Talking	.87	.62	.40	.858
Attention	.81	.54	.38	.645
Untidy	0	0	0	1.706
$\sigma$	.494	.483	.320	
$r_{III-V}$	.988			
$r_{III-80}$	.987			
$r_{V-80}$			.997	

Case III method for the stimuli employed in this experiment. This is true despite the fact that one assumption for Case V—approximate equality of the estimated discriminial dispersions—is grossly violated in each of the four samples.

2. Guilford's Short-Cut approach appears to yield essentially the same scale distribution as both the Case V and Case III methods for ordering the stimuli employed in this study. This is true despite the frequent observations in the literature that Guilford was unable to indicate a unit for his psychological scale.

In the opinion of the writer, these conclusions, taken in conjunction with the empirical findings of other investigators, and considered from the standpoint of contemporary rational developments, suggest a number of practical and theoretical implications. One possibility regarding the violation of the assumption of equal discriminial dispersions for Case V is indicated from Mosteller's work (22): He has reasoned that if an aberrant stimulus (i.e., dissimilar in discriminial dispersion) is near the center of the scale, there will not be much effect upon the ordering of the stimuli along the scale by means of Case V. This

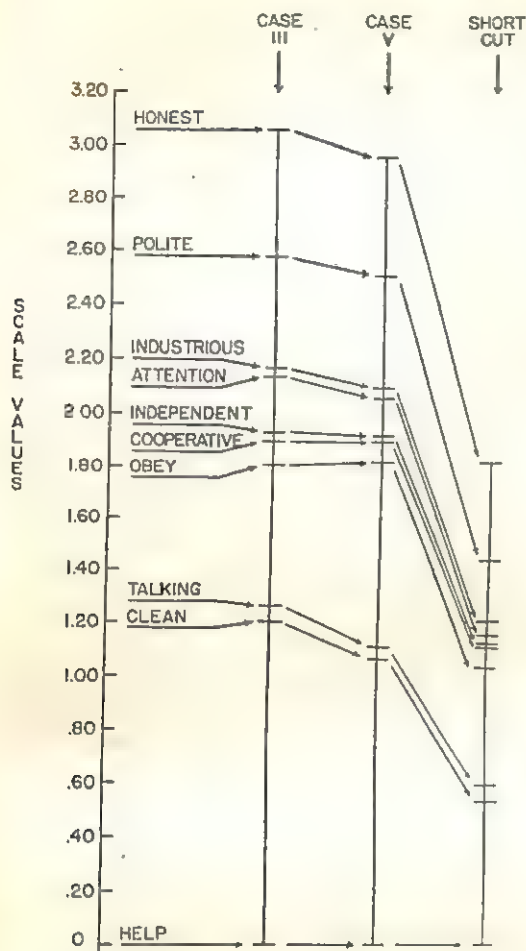


FIG. 2. "Teacher praise" scale values computed by Thurstone's Case III and Case V and Guilford's Short-Cut methods (80 twelfth-grade girls).

explanation provides a relaxation of the restriction that the largest discriminial dispersion should be no larger than twice the smallest dispersion for the employment of Case V (10). The most seriously aberrant stimuli in the data in the present investigation are for the most part the smaller ones, and they tend to fall near the middle of the scales in the four samples. It should also be kept in mind that Thurstone himself admitted that his calculated dispersions were *estimates* (28) and contained large probable errors.

An important practical consideration emerges from these possibilities. If Mosteller's rational efforts combined with the empirical findings in this study point toward an increasing generality of the applicability of

Case V, then the labor of calculations will be greatly reduced, as compared to Case III, and these conditions might then stimulate more widespread use of a very valuable, powerful, and somewhat neglected tool in psychological measurement, namely the method of paired comparisons. As a matter of fact, the classical reference to this tool as "psychophysical" is somewhat misleading since Thurstone has emphasized that (29, p. 142), "Although the law of comparative judgment is easily applied to the stimuli of classical psychophysics, the more generally interesting applications are those which involve social, moral, and esthetic values, opinion polls, and consumer preferences." More recently, the method of paired comparisons has been exploited in such diverse areas as sociometry (17, 31, 32), industry (18, 19, 25), social motivation (30), and learning theory (12, 33).

Guilford's Short-Cut method provides an even more economical approach than Thurstone's Case V. The shortcoming of Guilford's approach is the lack of an adequately defined psychological unit. Yet, it appears to "work," as demonstrated in the empirical findings reported in the present study. Perhaps a possible rationale for this approach

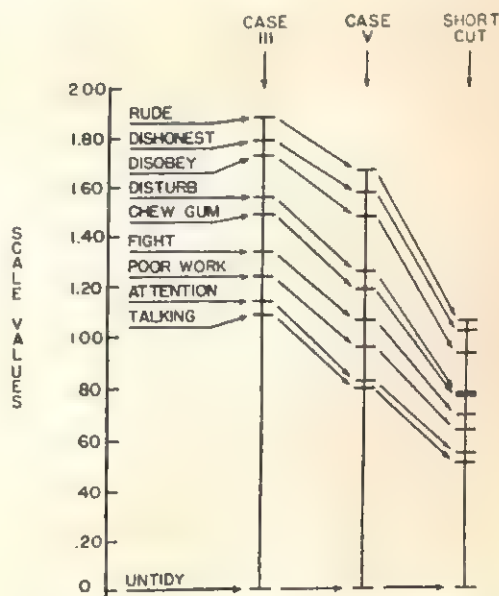


FIG. 3. "Teacher scold" scale values computed by Thurstone's Case III and Case V and Guilford's Short-Cut methods (80 eighth-grade girls).



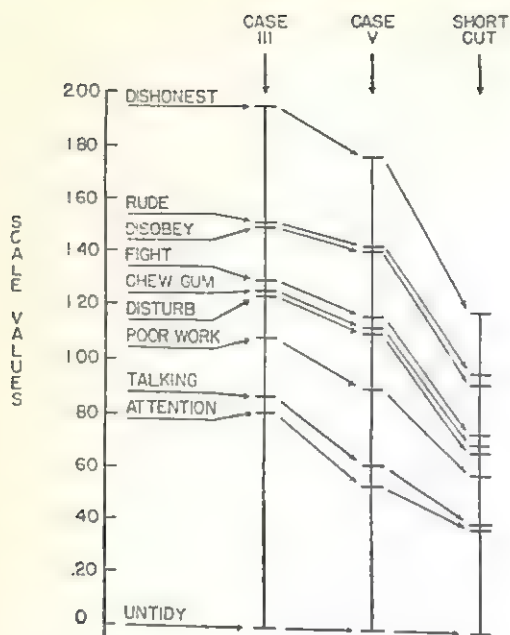


FIG. 4. "Teacher scold" scale values computed by Thurstone's Case III and Case V and Guilford's Short-Cut methods (80 tenth-grade boys).

might be found in Coombs' (5) "ordered metric" scaling concept, although he feels that the greater power of the method of paired comparisons is wasted, since the method of rank orders can be easily employed for his "psychological scaling without a unit of measurement." It is of interest to note here that Edwards (6, 7) has demonstrated the method of successive intervals to be an economical alternative to the method of paired comparisons.

If statistical theorists can continue to resolve some more of the rational problems of the method of paired comparisons, there is promise that this highly reliable technique will continue to be expanded to an increasingly larger number of qualitative problems in psychological measurement. This promise has been demonstrated by practical attempts to curtail the amount of labor involved in the subject's task, in the ordering of the pairs, and in scoring results. McCormick, Bachus, and Roberts (19, 20) studied the effects of decreasing the number of pairs upon the reliability of the resulting scales. Angoff (2) has investigated the problem of removing obsolete items from an established paired-

comparison scale and adding new items. Kephart and Oliver (16) introduced a punched card procedure as a laborsaving device for ordering pairs and scoring results. This combination of empirical research and rational development has fortified the usefulness of a powerful and extremely practical scaling technique, the method of paired comparisons. Psychologists interested in research with qualitative data will find a valuable aid here.

### Summary

The purpose of this study was to make an experimental comparison of Thurstone's Case III and Case V, and Guilford's Short-Cut approaches to scaling paired-comparison data, and to review recent rational and empirical developments of theoretical and practical significance for the application of paired comparisons to qualitative data. The stimuli were ten teacher-approved and ten teacher-disapproved behavior categories presented in paired-comparison form to four groups of school children. Each of the four groups contained a sample of 80 subjects and represented a particular sex, experimental condition (teacher-approved or disapproved behavior categories), and an age-grade level in the range from grades 6-12.

The intercorrelations between the scale values obtained by the three methods in the four samples for both sexes under both experimental conditions were approximately unity; twelve product-moment intercorrelations were .987 or higher. The results were interpreted as corroborative of recent rational and empirical investigations demonstrating the power of less complicated and economical approaches to scaling paired-comparison data than Thurstone's Case III, with the relaxation of certain restrictive assumptions. Possibilities for broader application of the method of paired comparisons to qualitative psychological problems were reviewed.

Received April 6, 1953.

### References

1. Andrews, T. G. Multidimensional psychophysics: a new research method. Paper read at Eastern Psychol. Ass., Atlantic City, March, 1952.

2. Angoff, W. H. An empirical approach to a problem of psychophysical scaling. *J. appl. Psychol.*, 1949, 33, 59-68.
3. Balinsky, B., Blum, M. L., and Dutka, S. The coefficient of agreement in determining product preferences. *J. appl. Psychol.*, 1951, 35, 348-351.
4. Burros, R. H. The application of the method of paired comparisons to the study of reaction potential. *Psychol. Rev.*, 1951, 58, 60-66.
5. Coombs, C. H. Psychological scaling without a unit of measurement. *Psychol. Rev.*, 1950, 57, 145-158.
6. Edwards, A. L. Psychological scaling by means of successive intervals. Univ. Chicago, Psychometric Laboratory Report No. 69, May, 1951.
7. Edwards, A. L. The scaling of stimuli by the method of successive intervals. *J. appl. Psychol.*, 1952, 36, 118-122.
8. Guilford, J. P. The method of paired comparisons as a psychometric method. *Psychol. Rev.*, 1928, 35, 494-506.
9. Guilford, J. P. Some empirical tests of the method of paired comparisons. *J. gen. Psychol.*, 1931, 5, 64-77.
10. Guilford, J. P. *Psychometric methods*. New York: McGraw-Hill, 1936.
11. Gulliksen, H. P. Paired comparisons and the logic of measurement. *Psychol. Rev.*, 1946, 53, 199-213.
12. Hull, C. L., Felsing, J. M., Gladstone, A. I., and Yamaguchi, H. G. A proposed quantification of habit strength. *Psychol. Rev.*, 1947, 54, 237-254.
13. Johnson, P. O. *Statistical methods in research*. New York: Prentice-Hall, 1949.
14. Kendall, M. G. *The advanced theory of statistics*. London: Charles Griffin & Co., 1945.
15. Kendall, M. G. and Smith, B. B. On the method of paired comparisons. *Biometrika*, 1940, 31, 324-345.
16. Kephart, N. C. and Oliver, J. E. A punched card procedure for use with the method of paired comparisons. *J. appl. Psychol.*, 1952, 36, 47-48.
17. Koch, Helen L. A study of some factors conditioning the social distance between the sexes. *J. soc. Psychol.*, 1944, 20, 79-107.
18. Lawshe, C. H., Kephart, N. C., and McCormick, E. J. The paired comparison technique for rating performance of industrial employees. *J. appl. Psychol.*, 1949, 33, 69-77.
19. McCormick, E. J. and Bachus, J. A. Paired comparison ratings. I. The effect on ratings of reductions in the number of pairs. *J. appl. Psychol.*, 1952, 36, 123-127.
20. McCormick, E. J. and Roberts, W. K. Paired comparison ratings. II. The reliability of ratings based on partial pairings. *J. appl. Psychol.*, 1952, 36, 188-192.
21. Mosteller, F. Remarks on the method of paired comparisons: I. The least squares solution assuming equal standard deviations and equal correlations. *Psychometrika*, 1951, 16, 3-9.
22. Mosteller, F. Remarks on the method of paired comparisons: II. The effect of an aberrant standard deviation when equal standard deviations and equal correlations are assumed. *Psychometrika*, 1951, 16, 203-206.
23. Mosteller, F. Remarks on the method of paired comparisons: III. A test of significance for paired comparisons when equal standard deviations and equal correlations are assumed. *Psychometrika*, 1951, 16, 207-218.
24. Mosteller, F. Some miscellaneous contributions to scale theory: Remarks on the method of paired comparisons. Harvard University Laboratory of Social Relations, Report No. 10, 1952.
25. Satter, G. A. Method of paired comparisons and a specification scoring key in the evaluation of jobs. *J. appl. Psychol.*, 1949, 33, 212-221.
26. Thurstone, L. L. A law of comparative judgment. *Psychol. Rev.*, 1927, 34, 273-286.
27. Thurstone, L. L. The measurement of opinion. *J. abnorm. soc. Psychol.*, 1928, 22, 415-430.
28. Thurstone, L. L. Stimulus dispersions in the method of constant stimuli. *J. exp. Psychol.*, 1932, 15, 284-297.
29. Thurstone, L. L. Psychophysical methods. In Andrews, T. G. (Ed.), *Methods of psychology*. New York: Wiley, 1948, Chapt. 5, 124-157.
30. Witryol, S. L. Age trends in children's evaluation of teacher-approved and teacher-disapproved behavior. *Genet. Psychol. Monogr.*, 1950, 41, 271-326.
31. Witryol, S. L. and Thompson, G. G. A critical review of the stability of social acceptability scores obtained with the partial-rank-order and the paired-comparison scales. *Genet. Psychol. Monogr.*, in press.
32. Witryol, S. L. and Thompson, G. G. An experimental comparison of the stability of social acceptability scores obtained with the partial-rank-order and the paired-comparison scales. *J. educ. Psychol.*, 1953, 44, 20-30.
33. Zeaman, D. An application of  $sE_R$  quantification procedure. *Psychol. Rev.*, 1949, 56, 341-350.



## Reliability and the Number of Rating Scale Categories

A. W. Bendig<sup>1</sup>

*University of Pittsburgh*

A recent study (1) has presented evidence that the reliability of rating scales is independent of the number of categories on the scale. In this study Ss rated themselves on their comparative knowledge about 12 foreign countries. The scales used varied in the number of verbal anchors used to define the scale categories (1, 2, or 3) and in the number of categories (3, 5, 7, 9, or 11). Both group and individual rater reliability was relatively invariant over the range from three to nine categories with a slight drop in reliability at eleven scale points. These results contradict the theoretical analysis of Symonds (7) who concluded that scale reliability should increase with greater numbers of scale categories, but that this increase becomes negligible above nine scale points.

However, the empirical results reported were germane only to one type of reliability: the reliability with which Ss can distinguish between stimuli presented to them. This is the type of reliability analysis that is important when raters are given the task of rating stimuli on some criterion scale and the mean rating for each stimulus is to be used as the criterion measure. The assessment of the reliability of pooled supervisor ratings of workers is a practical example of this type of problem. A second question subject to reliability analysis is how well a series of self-ratings discriminates among the Ss. Many psychometric instruments can be regarded as a series of stimuli presented to the Ss with the request that the S rate himself on a two, three, or R category scale. For example, the Strong Vocational Interest Blank commonly requires self-rating on a three-category scale while the revised Bogardus Scale of Social Distance uses a seven-point scale. Commonly the total score of an S on these instruments is the sum or mean of his ratings and the reliability question concerns the

ability of the test's total score to discriminate among the Ss. This type of reliability is the more usual "test reliability" compared to the first type which we might call "rater reliability." Our first study (1) suggested that Symonds' analysis does not hold for rater reliability, but did not present evidence concerning test reliability.

The present report concerns a study of the reliability of food preference ratings. Some years ago Wallen (9) found that responses to a check list of food aversions significantly discriminated between groups of normal and neurotic military personnel. Because of the restricted range of food aversions in his normal groups Wallen reports no reliabilities for normal Ss. However, the data given (9, pp. 79-80) permit the application of Kuder-Richardson formula 20 (8, p. 92). Using this formula the reliabilities for two normal groups are estimated to be .28 ( $N = 100$ ) and .82 ( $N = 114$ ). The weighted mean ( $r$ -to- $z$  transformation) of these two estimates is .64. The original Wallen check list used a rating scale with only two categories: strong dislike or acceptance of the 20 food stimuli presented to the S. Symonds' conclusions would suggest that the somewhat low reliability of this instrument should increase if the Ss were allowed to rate the foods on rating scales containing more categories that would permit the Ss to make finer discriminations among the foods.

### Procedure

**Scales.** The stimuli to be rated by the Ss were the list of 20 foods used by Wallen (9); This list was given to each S with a rating scale having either 2, 3, 5, 7, or 9 categories. For the two-category scale the instructions used by Wallen (9, p. 78) were given. These instructions were modified for the other scales and three anchoring statements were used to describe the center and end categories on these scales. The statements used were: (a)

<sup>1</sup> Miss Janine Sprague assisted with some of the statistical computations.

I like this food very much and eat it often; (b) I am somewhat neutral toward this food, neither liking nor disliking it much; and (c) I dislike this food so much that I refuse to eat it.

The anchored scales with unit digits (1, 2, etc.) designating the categories were mimeographed on single sheets with the list of 20 foods and randomly distributed to the Ss.

*Subjects.* The Ss were 249 students in introductory and social psychology classes. The ratings of Ss were excluded from the analysis whenever the S used less than one-half of the available categories in rating the foods. Thus an S, using the five-category scale, who used only ratings of one and five was not included. A total of 13 Ss was eliminated under this criterion, giving a study group of 236 Ss.

*Analysis.* Test reliability was assessed using the analysis of variance technique devised by Hoyt (4, 5) and recommended by Thorndike (8, pp. 93-96). Rater reliability was estimated by the similar procedures described by Ebel (2). Since the number of raters varied slightly from scale to scale, the reliability of a single rater was computed to adjust for this varying N. Confidence limits (90 per cent) were computed following Ebel (2, pp. 413-414). Finally, the average rank-difference correlations between the rankings of the foods on the five rating scales were computed (6, pp. 80-84).

### Results

The test reliability and rater reliability estimates for each rating scale can be found in Table 1 along with the 90 per cent confidence interval for each reliability. The five test reliability estimates were tested for homogeneity using the chi-square method described by Edwards (3, p. 135). The resulting chi-square value was 1.12 which, with four degrees of freedom, is not significant at the .05 confidence point. The mean reliability was .625. A similar test of the homogeneity of the rater reliabilities gave a chi-square of 1.76 which again is not significant. The mean rater reliability was 0.23.

The average rank-difference correlation between the rankings of the 20 foods on the five scales was .90 when corrected for ties. Since

Table 1

Reliability Estimates of Food Preference Rating Scales with Various Numbers of Scale Categories

	Number of Rating Scale Categories				
	2	3	5	7	9
Number of Subjects	52	41	52	46	45
Test Reliability	.61	.63	.58	.70	.60
Confidence Limits (.90)					
Upper	.72	.74	.69	.78	.71
Lower	.44	.43	.39	.56	.40
Rater Reliability	.07	.33	.25	.24	.24
Confidence Limits (.90)					
Upper	.12	.45	.35	.34	.34
Lower	.03	.20	.14	.13	.13

there were a number of foods tied in rank on the scales with two and three categories a similar average rho was computed on the food rankings on scales with five, seven, and nine categories and was found to be .91.

### Discussion

The results in terms of the test reliabilities is fairly unequivocal. No consistent trend was found in the relation of test reliability and number of scale categories. This suggests that Symonds' (7) analysis does not hold for test reliability. It is interesting to note that the mean reliability found for the five scales, .625, is very similar to the estimate from Wallen's data, .64. While the highest reliability, .70, was found with seven categories, the two lowest reliabilities, .58 and .60, were found with the immediately adjacent numbers of categories (five and nine).

Rater reliability was not as regular as test reliability. The invariance of reliability over the range of five to nine categories that was found in a previous study (1) is here confirmed. However, rater reliability rose at three categories and dropped for two categories in this study. The drop at two may be attributable to the slightly different instructions to the Ss with this scale. The slightly greater reliability with three categories cannot be explained by different instructions, although it must be pointed out that this reliability is not much higher and, when tested



statistically, is not significant. Before we can extend the conclusion of invariant reliability below five scale categories further investigation will be necessary.

It is interesting to note that the rater reliabilities found for our list of 20 foods is somewhat less than that found for ratings of foreign countries (1, p. 39). This lower rater reliability for foods may be a function of the type of judgment required of the Ss, of the greater number of judgments required of the Ss (20 instead of 12), or of a greater homogeneity among the 20 foods than was present among the 12 countries.

### Summary

Ss ( $N = 236$ ) rated 20 foods as to preference using rating scales containing 2, 3, 5, 7, and 9 categories. Test reliability (summed ratings for each S) and rater reliability (summed ratings for each food) were computed for each scale. Test reliability was constant over the entire range of categories and was very similar to reliabilities found in another study. Rater reliability was constant from five to nine categories, but was slightly lower at two and slightly higher at three categories. It was concluded that test reliability is independent of the number of

scale categories, and that rater reliability is relatively constant, but that further research on rater reliability using short scales is needed before a similar generalization can be made regarding rater reliability.

Received April 18, 1953.

### References

1. Bendig, A. W. The reliability of self-ratings as a function of the amount of verbal anchoring and of the number of categories on the scale. *J. appl. Psychol.*, 1953, 37, 38-41.
2. Ebel, R. L. Estimation of the reliability of ratings. *Psychometrika*, 1951, 16, 407-424.
3. Edwards, A. L. *Experimental design in psychological research*. New York: Rinehart, 1950.
4. Hoyt, C. Test reliability obtained by analysis of variance. *Psychometrika*, 1941, 6, 153-160.
5. Hoyt, C. J. and Stunkard, C. L. Estimation of test reliability for unrestricted item scoring methods. *Educ. psychol. Measmt*, 1952, 12, 756-758.
6. Kendall, M. G. *Rank correlation methods*. London: Griffin, 1948.
7. Symonds, P. M. On the loss of reliability in ratings due to coarseness of the scale. *J. exp. Psychol.*, 1924, 7, 456-461.
8. Thorndike, R. L. *Personnel selection*. New York: Wiley, 1949.
9. Wallen, R. Food aversions of normal and neurotic males. *J. abn. soc. Psychol.*, 1945, 40, 77-81.

# The Inference of Accident Liability from the Accident Record

Alexander Mintz

*City College of New York*

It has been known for a long time that accident liability<sup>1</sup> of people, that is their potential long range accident rate, and the actual number of accidents occurring to them, i.e., their accident record, are imperfectly correlated. This was already clearly implied in the classical 1920 paper by Greenwood and Yule (3) on accidents. Newbold (9) presented in 1927 a formula for estimating the correlation between accident records and accident liability. Cobb (1) pointed out in 1940 that this correlation need not be high. Mintz and Blum (8) examined a large number of published distributions and found that the estimated variance of accident liability usually accounts only for a relatively small portion of the variance of accident records, thus confirming Cobb's finding. Quite recently, Hughes (4) included in his summary of the mathematical research on accidents tables and graphs implying the imperfect correlation between accident liability and records. These tables and graphs utilize Greenwood and Yule's theoretical inference that for any one particular degree of liability in people there should be a Poissonian distribution of accident records. His table presents the probability, for different degrees of accident liability and for different mean group liabilities, that a person should have twice as many accidents as the mean for the total group.

In all papers mentioned the notion is utilized that a given degree of accident liability tends to result in a Poisson distribution of accident records. This notion has many theoretical uses, but its practical usefulness is limited by the fact that in the case of particular individuals the degree of accident liability is generally unknown, so that the Poissonian probability distributions of accident records

cannot be arrived at. The accident record of individuals is often available.<sup>2</sup> What is often needed is a procedure for estimating the unknown accident liability in terms of the known accident record. The main problem of this paper is: given a known distribution of accident records, and a particular accident record belonging to this distribution, how probable are the different assumed degrees of accident liability which may correspond to this particular accident record?

In this general form, the problem has no answer. It will be treated here in terms of certain assumptions first explored theoretically by Greenwood and Yule (3) and empirically by Greenwood and Woods (2). These assumptions were:

- (1) accident liability of people is not changed by accidents in which they are involved and does not vary with time;<sup>3</sup> and
- (2) accident liability varies among people and is distributed in some known manner, e.g., in accordance with a Pearson Type III curve.

These assumptions have not been definitely shown to be true, but they are fairly well supported by available evidence, so that a further exploration of their implications is in order.

## The Solution

The following considerations indicate the nature of the solution. Accident liability and accident record may be treated as two correlated variables, the former as the independent, the latter as the dependent variable. The distribution of accident liability is assumed to be known, or to be capable of being estimated from the data; so are the theoretically Poissonian distributions of accident records in the columns of the scatter diagram. To-

<sup>1</sup> "Accident liability" is a more general term than accident proneness because it includes both personal and environmental conditions predisposing people to accidents. Exact constancy of environmental hazards is hard to prove, so that it is probably normally more accurate to refer to accident liability rather than proneness.

<sup>2</sup> Or approximately known. There are a number of pitfalls in the way of precise characterization of accident records which have been discussed in the literature.

<sup>3</sup> Actually, a somewhat weaker assumption is sufficient, as has been pointed out by Kerrich (6).



gether these two types of information define a complete correlation surface; this correlation surface describes the probability distribution of various possible combinations of accident liability and accident record. There should be no difficulty in determining the distributions in the rows of such a correlation surface. Such a distribution would indicate how probable are various degrees of accident liability in the case of a particular accident record, presupposing the assumptions of accident liability being unaffected by the occurrence of accidents and having a known distribution.

The mathematical derivation of such a distribution is presented below. It assumes as was suggested by Greenwood and Yule that accident liability is distributed in a Pearson Type III curve. In this particular case the answer is a very simple one: If the distribution of accident liability for the whole group is of the Pearson III type, then the probable distributions of accident liability are also of the Pearson III type, but with changed constants in the formula. The changing of the constants results in changed means and standard deviations which vary from those of the whole group, and also vary according to the accident record of the specific subgroups.

### Mathematical Derivation

*Poisson distribution:* Probability of  $j$  accidents for group with accident liability  $\lambda$ :

$$\frac{e^{-\lambda} \lambda^j}{j!},$$

where  $e = 2.718 \dots$  (base of natural logarithms).

*Pearson III distribution of accident liability:*

$$\frac{c^p}{\Gamma(p)} e^{-c\lambda} \lambda^{p-1},$$

where  $\lambda$  is liability and  $c$  and  $p$  are constants related as follows to the mean ( $m$ ) and variance

( $v$ ) of the distribution:  $m = \frac{p}{c}$ ,  $v = \frac{p}{c^2}$ .

*Greenwood-Yule derivation of negative binomial distribution:* Probability of  $\lambda$  liability and  $j$  accidents: product of formulae for Pearson III

and Poisson distributions:

$$\frac{c^p}{\Gamma(p)} e^{-c\lambda} \lambda^{p-1} \times \frac{e^{-\lambda} \lambda^j}{j!} = \frac{c^p}{j! \Gamma(p)} e^{-(c+1)\lambda} \lambda^{p+j-1}.$$

To determine the probability of  $j$  accidents for all  $\lambda$ -s, this expression has to be integrated over all values of  $\lambda$ , so that  $0 \leq \lambda < \infty$ .

$$\begin{aligned} \int_0^\infty \frac{c^p}{j! \Gamma(p)} e^{-(c+1)\lambda} \lambda^{p+j-1} d\lambda &= \\ \left( \text{if } (c+1)\lambda = x, d\lambda = \frac{dx}{c+1} \right) &= \\ = \frac{c^p}{j! \Gamma(p)} \int_0^\infty \frac{e^{-x} x^{p+j-1}}{(c+1)^{p+j-1} c+1} dx &= \\ = \left( \frac{c}{c+1} \right)^p \frac{\int_0^\infty e^{-x} x^{p+j-1} dx}{j! \Gamma(p) (c+1)^j} &= \\ = (\text{by definition of } \Gamma\text{-function}) &= \\ = \left( \frac{c}{c+1} \right)^p \frac{\Gamma(p+j)}{j! \Gamma(p) (c+1)^j} \end{aligned}$$

(general term of negative binomial distribution).

*Derivation of probable distribution of accident liability for given accident record:* Probability of accident liability  $\lambda$  and accident record  $j$ , in relation to all possible combinations of  $\lambda - s$  and  $j - s$ :

$$\frac{c^p}{j! \Gamma(p)} e^{-(c+1)\lambda} \lambda^{p+j-1}.$$

Probability of accident liability  $\lambda$  and accident record  $j$ , in relation to the combined probability of combinations of this particular  $j$  with all  $\lambda - s$ :

$$\begin{aligned} \frac{c^p}{j! \Gamma(p)} e^{-(c+1)\lambda} \lambda^{p+j-1} &= \\ \left( \frac{c}{c+1} \right)^p \frac{\Gamma(p+j)}{j! \Gamma(p) (c+1)^j} &= \\ = \frac{c^{p+j}}{\Gamma(p+j)} e^{-(c+1)\lambda} \lambda^{p+j-1}. \end{aligned}$$

(Estimated distribution of accident liability  $\lambda$  corresponding to given accident record  $j$ . The distribution is one of Pearson's Type III. It has an equation of the same form as that given above for the Pearson III distribution, but with changed constants;  $c+1$  replaces  $c$ ,  $p+j$  replaces  $p$ .)

The estimated Pearson III curves of accident liability for subgroups with given accident records may be interpreted in two ways: first, as representing the probable numbers of people with various levels of accident liability in a subgroup with a given accident record; and, second, as representing the degrees of probability of these various levels of liability for people with given accident records. Only the second interpretation is appropriate in the case of small subgroups.

### Illustrative Results

The accident distribution reported in Greenwood and Woods' Table 8A was used for the computation of Pearson III curves as just explained. This set of data was chosen for two reasons: (1) it can be closely approximated by the theoretical distribution derived from the Greenwood-Yule assumptions (the so-called negative binomial distribution), which suggests that these assumptions may hold true in this case; (2) this set of data was suggestive of a higher correlation between accident record and liability than the other Greenwood and Woods sets of data (2). It was thought therefore that the demonstration of a relatively wide spread of probable accident liability corresponding to a particular accident record would be particularly convincing. Table 1 presents this set of data, together with the negative binomial distribution fitted by the method of moments.

Figure 1 presents the three Pearson III

Table 1

Accid.	No. of People	Theoretical (Neg. Binomial)
0	8	8.7
1	11	10.1
2	8	8.9
3	10	6.9
4	3	5.0
5	4	3.5
6	1	2.4
7	—	1.6
8	2	1.0
9	2	0.7
10	—	0.4
11	1	0.3
Total	50	49.5

curves for the subgroups with zero, five, and eleven accidents (a subgroup of one).

The curves show that a large group whose members have had five accidents apiece is likely to include some persons whose potential accident records have a very wide range. There is actually some noticeable overlapping even between the probability curves of liability of the two extreme subgroups with zero and eleven accidents.

There is a very considerable amount of overlapping between the liability curve for the five-accident group and the other two.

The estimated Pearson III distributions of accident liability for people with given accident records enable one to estimate the combined probability of their accident liability falling within certain ranges, e.g., the range below the mean of the whole group or above twice the group mean, or from the first to the third quartile of the whole group. One can do this by integrating the expression for the Pearson III curve, or by using tables of the Pearson III integral (e.g. 10).

Table 2 presents the results of such a procedure for two published distributions of accidents. These distributions are that in Greenwood and Woods' Table 8A and that of 29,531 Connecticut car drivers discussed by Cobb (1). The figures represent the probability that the true accident liability of a person with a given accident record is below the mean<sup>4</sup> of the whole group. The two means were 2.8 accidents per person and .24 accidents per person for the Greenwood-Woods set 8A and for the Connecticut drivers, respectively.

In the Greenwood-Woods' set of data, a person who had no accidents has 95.2 chances in a hundred of having accident liability below the mean of the whole group. For people with 1 accident, the probability of accident liability below the group mean of 2.8 is 85.7 per cent, and so on. Similar statements can be made about the Connecticut car drivers. It should be noted that 95 per cent

<sup>4</sup> In the subsequent discussions, there are references to the probability of accident liability being either above or below the group mean. This is done for the sake of simplicity; the infinitesimal probability of accident liability being exactly at the group mean is disregarded.



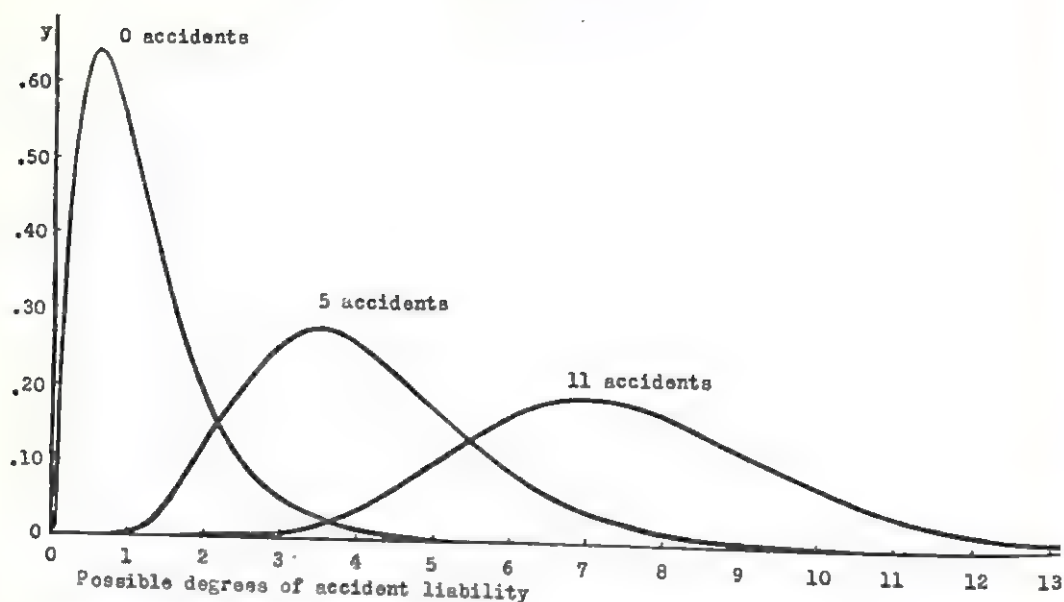


FIG. 1. Pearson Type III curves ( $y = \frac{c^{p+j}}{\Gamma(p+j)} e^{-(c+1)\lambda} \lambda^{p+j-1}$ ) representing the probability distributions of different degrees of accident liability for given accident records. Source of data: Greenwood and Woods, Table 8A.

certainty about accident liability being greater than that of the group mean is not reached until the person has had 8 or more accidents in the Greenwood-Woods' sample, 3 or more accidents in the sample of Connecticut drivers. These accident records, are reached by only a few persons—5 out of 50, or 10 per cent in the former case, 51 out of 29,531 or .17 per cent in the latter case.

Table 2

Probability (in Per Cent) of Accident Liability Below the Mean of the Whole Group

Number of Accidents	Greenwood-Woods Table 8A	Connecticut Drivers
0	95.2	76.6
1	85.7	40.9
2	70.6	15.9
3	52.6	4.9
4	33.0	1.1
5	20.9	0.3
6	11.3	0.1
7	5.6	
8	2.5	
9	1.1	
10	0.4	
11	0.1	

### Discussion

The immediately preceding statements should not be confused with the customary statements of the level of statistical significance. If one states that a finding is significant at the 5 per cent level, one means that, if the null hypothesis is assumed to be valid for the population, deviations as great or greater than the one found are expected to be found in only 5 per cent of the samples. The statement, "the probability of accident liability below the group mean is 5 per cent in people having 3 accidents" does not presuppose an assumed null hypothesis and is not intended to be a test of a null hypothesis. On the contrary, it presupposes the existence of differences in accident liability and characterizes the probable percentage of below-average liability among people who had 3 accidents each.

Taken at their face value, the figures in Table 2 exhibit the way in which accident liability below the group mean becomes less probable and accident liability above the mean becomes more probable in the case of persons with the larger numbers of accidents. Clearly, the accident records have some va-

liability as information about accident liability. In the Greenwood-Woods set of data, 5 or more accidents mean accident liability above the group mean in at least four cases out of five. In the Connecticut drivers' sample, drivers who had 2 or more accidents can be expected to have accident liability above the group mean in more than 5 cases out of 6. On the other hand, relative certainty that a given individual has accident liability above the mean of the whole group can only be achieved in a very small number of people. Whether one chooses to emphasize the fact that accident records have some validity as long range predictors, or the limitations of their validity is presumably dependent on one's scientific level of aspiration.

Should figures be taken at their face value? The answer depends on whether the Greenwood-Yule unequal liability assumptions are to be accepted. The principal evidence on their validity seems to be as follows:

1. The negative binomial distribution which, as theoretically derived by Greenwood and Yule, was based in part on these assumptions fits many obtained accident distributions very well. Newbold (9) showed, in effect, that it usually fits them better than any other theoretical distribution embodying the ideas of unequal accident liability in people and unchanged accident liability after accidents.

2. The negative binomial distribution can be derived by the use of different sets of assumptions and therefore does not differentiate between them and the Greenwood-Yule assumptions. Thus Irwin (5) showed that the negative binomial distribution is to be expected if there are no initial differences in accident liability and that accident liability of people increases as a linear function of accidents. Lundberg (7) quotes rather similar deductions by Polya and Eggenberger.

3. There are a few available sets of accident data for the same people during consecutive periods. In terms of the Greenwood-Yule assumptions the accident rate of people should remain constant. In terms of the assumptions explored by Irwin they should increase. According to the evidence presented by Irwin and by Kerrich (6), the accident

rates vary only slightly with time, and tend to decrease rather than to increase.

In terms of the evidence presented, the major inferences from the Greenwood-Yule assumptions appear to be in accord with available data in many cases. However, more research is needed, particularly in view of the scanty available evidence on accidents in successive periods. There are theoretical considerations making the exact truth of Greenwood and Yule's assumptions of unchanged liability after accidents rather unlikely. Nevertheless, the available evidence strongly suggests that in the cases in which the negative binomial distribution fits the data the Greenwood-Yule assumptions may be viewed as approximating the truth. The inferences from these assumptions pertaining to the probable degree of accident liability which may correspond to given accident records then may be tentatively accepted as approximately true in many cases.

#### Summary

The classical assumptions of unchanged accident liability after the occurrence of an accident were provisionally accepted. Certain further implications of these assumptions were explored. The assumed distributions of accident liability in groups of people were broken up into probable component distributions of liability for subgroups with given accident records. These component distributions were found to have the same form as the total distribution if the latter is of type III. Quantitative examples of applications of this finding were given. It was pointed out that accident records have some validity as indicators of accident liability, but that relative certainty about high accident liability of particular persons can be achieved in terms of their accident records only in a small minority of cases.

*Received April 20, 1953.*

#### References

1. Cobb, P. W. The limit of usefulness of accident rate as a measure of accident proneness. *J. appl. Psychol.*, 1940, 24, 154-159.



2. Greenwood, M. and Woods, H. M. The incidence of industrial accidents upon individuals with specific reference to multiple accidents. *Industr. Fatigue Res. Bd.* Report 4, 1919.
3. Greenwood, M. and Yule, G. U. An enquiry into the nature of frequency distributions representative of multiple happenings, with particular reference to the occurrence of multiple attacks of disease or of repeated accidents. *J. Roy. Statist. Soc.*, 1920, 83, 255-279.
4. Hughes, H. M. Discriminatory analysis. III. Discrimination of the accident prone individual. *USAF School of Aviation Medicine*. Project Report number 21-49-004. Report number 3. Oct. 1950.
5. Irwin, J. O. Comments on paper by Chambers, E. G. and Yule, G. U. *J. Roy. Statist. Soc.*, 1941, 7 (suppl.), 101-107.
6. Kerrich, J. E. The mathematical background. In Arbous, A. G. and Kerrich, J. E. *Accident statistics and the concept of accident proneness*. National Institute for Personnel Research, No. 391, 1951. South African Council for Scientific and Industrial Research.
7. Lundberg, O. *On random processes and their application to sickness and accident statistics*. Uppsala, Almqvist and Wiksell, 1940.
8. Mintz, A. and Blum, M. L. A re-examination of the accident proneness concept. *J. appl. Psychol.*, 1949, 33, 195-211.
9. Newbold, E. M. Practical applications of the statistics of repeated events. *J. Roy. Statist. Soc.*, 1927, 90, 487-547.
10. Pearson, K. (editor). *Tables of the incomplete Gamma function*. London, H. Maj. Station: Off., 1922.

# The Development of Criteria of Safe Operation for Groups<sup>1</sup>

Harry Waller Daniels and Harold A. Edgerton

*Richardson, Bellows, Henry & Co., Inc., 1 West 57 Street, New York 19, N. Y.*

It is often a task of the psychologist, in investigations of relationships between various aspects of the operation or management of groups and the effectiveness of the groups, to find some way of measuring this effectiveness. When group effectiveness is the goal, these attempts at measurement can be extremely difficult. Many times they fall short of the mark and do not really measure what the managers of the groups would themselves call effectiveness.

In a study of the relationship of various management factors to the relative safety effectiveness of Army motor vehicle units, done under contract with the Department of the Army, the present authors and their colleagues at Richardson, Bellows, Henry and Company came to grips with this knotty criterion problem.

The primary objective of the study was to determine the relationship, if any, between the driving safety of motor vehicle units, and management and supervisory practices used in those units. The orientation of the present paper is limited to a summary of the attempts to define and measure the criterion variables.

Although few psychologists have done intensive work in this area, the history of research on accident proneness, safe driving, and traffic problems is long and varied. Two excellent reviews are available: Johnson's (1) and Lawshe's (2). The authors of these reviews show that few investigators have differentiated between driving *skill* and *safe* driving, and that the early investigators concentrated on "simpler" functions like depth perception rather than "higher" functions like abilities, attitudes, etc.

The research presented in this paper has as its orientation the *safe operation* of a motor vehicle *unit*, rather than the more usual orientation of skill in driving. Thus, the first

question to which efforts were directed was: How could the units which were "high" and those which were "low" in safety of operation be properly identified so that another observer using the same procedures would arrive at the same identification?

Preliminary investigation of accident rates of motor vehicle units led to the conclusion that, for the purposes of this study, such data were inadequate. Differences in definition of a reportable accident, in accuracy of mileage estimates, in mission, in equipment, traffic conditions encountered, etc., were factors involved. Added to this was the statistical unreliability of reported accident rates for, say, 50 vehicle motor units.

Such information and evidence led toward ratings of safety of performance as a method of identifying units for further study. In addition, it was not feasible to restrict the study to motor vehicle units which were fairly comparable in organization, equipment, mission, and conditions of operation. To produce really useful results, the study had to encompass motor units as they occurred rather than as one might like to have them set up for a "tight" experimental design.

## Procedure

The forms used were constructed on the basis of the preliminary surveys and the field tryout. The criterion procedure was as follows:

Relative ratings of over-all safety of operations of all motor vehicle units in an installation were asked for. Criterion rating sessions were held, attended by post or divisional staff officers, Provost Marshals, Safety Officers and Directors, and other persons who would have an acquaintance with the comparative performances of the motor vehicle units at the installation. Motor officers and the sergeants of the individual motor units did not attend these sessions but were asked to fill out the rating forms on their own units at the time of intensive study (not reported here) of their own units.

In the criterion rating sessions three forms were filled out by the participants after pertinent

<sup>1</sup> The opinions and conclusions expressed in this article are those of the authors; they do not necessarily reflect official Department of the Army policy or the views of anyone other than the authors.



instructions. Since these meetings were informal gatherings, it was possible for the RBH field representative to monitor the rating procedures of the participants and thus ensure that the instructions were being followed. Discussion of units being rated was not permitted; but otherwise conversation flowed freely during these sessions. The forms used were as follows:

The *Familiarity Rating Form* (CRT-29, RBH Form R212-J) was constructed for the raters to indicate how well they knew each of the motor vehicle units at a particular installation. This form was designed to overcome the objections frequently made by those officers that they could not judge the over-all safety of a motor unit because they were not sufficiently familiar with it. The form was used to identify the motor vehicle units with which they were best acquainted.

While the use of the *Familiarity Rating Form* did not completely overcome their reluctance, it seemed to relieve some tension in the criterion rating sessions. The form merely listed the motor units in the post or division, and the officers were instructed to rate their familiarity with the units as follows:

"0"—if unfamiliar

"1"—if slightly familiar

"2"—if familiar with the unit's personnel, driving, and other factors to be rated.

The *Safety Factors Rating* (CRT-26, RBH Form R212-G) was then given to the group members who were asked to rate, on 16 aspects of over-all safety, the six units with which they were most familiar. The raters who did not know six units well enough to rate them rated only those with which they had indicated familiarity.

The *Criterion Ranking Form* (CRT-27, RBH Form R212-H) was then administered. On this form the raters identified in order, from those with which they were familiar, up to six units which they thought were "best" from a viewpoint of all-around safety, and up to six units which were "worst" in all-around safety.

From the analysis of these forms it was possible to select a number of "high" safety and "low" safety units from each post or division. In many cases, upon further acquaintance with the unit (e.g., the Xth Ordnance Battalion), it was found that the unit which was selected as high or low really contained more than one motor unit (e.g., Companies A, B, C, and Hq.). In such cases, the Battalion, Regimental, or Group staff went through the criterion procedures as outlined above for the motor units under their cognizance. Company or Battery level motor units were selected from these ratings, with this limitation: only the units rated lowest were selected from Battalions, Regiments, or Groups previously rated low, and only the highest were selected from units previously rated high.

## A Check on the Criterion Groups

In spite of the experience gained early in the study at various Army installations, which showed that the usual accident and mileage records were unsuitable for our purposes, an attempt was made to provide an objective criterion measure of this general type.

It had originally been planned to collect accident frequency statistics for the units. Review of unit safety records during the pilot field study had indicated that accident frequency statistics were inadequate to permit differentiation among relatively safe and unsafe motor units. The results of this trial study, however, showed that many incidents, which could be construed as related to the safe operation of motor units, were not being reported as accidents. To utilize this information, the *Vehicle Damage Report* (CRT-30, RBH Form R212-K)<sup>2</sup> was devised. It was hoped that this form would provide a higher degree of objectivity and serve to substantiate or refute the selection of the units on the basis of the ratings and rankings. Data on the frequency of damages occurring within individual units, as an empirical measure of their safety of operation, were collected. The report was essentially a list of approximately 50 damages which could occur to a vehicle as a result of an impact. These were compiled and divided into nine general areas (e.g., Bumper Assembly, Body-Front, Body-Sides, Wheel Assembly, etc.). The list was further subdivided into types of vehicles.

The list was administered as a group interview with the units' motor sergeants, motor officers, mechanics, etc. Copies of the list were handed to each member of a group so that the list could serve as a stimulus to recognition and recall of damages which had occurred to the unit's vehicles during the preceding calendar month. They were encouraged to use whatever records they had available, and to consider each vehicle separately, one at a time.

## Analysis and Final Criterion Groups

From two division and four post headquarters, 93 motor units were rated by varying numbers of raters. From these 93, 16 "high" units and 16 "low" units were chosen. The analysis of these criterion measures and how they were used in making the choices are given below.

The scoring for the *Safety Factors Rating* was based on results of the preliminary field study at different installations in which an

<sup>2</sup> The authors are indebted to Mr. Warren R. Graham of Richardson, Bellows, Henry and Company for the development of the *Vehicle Damage Report*.

empirical key was derived and shown to be related to the ranking of motor units ( $r = .46$ ). Using all of the 93 rated units from the present sample, this empirical scoring was shown to be related to the rankings,  $r = .83$ . This high relationship may reflect considerable "halo," but as far as one can rely on the validity of the ratings in the criterion rating sessions, this consistency serves to substantiate the identification of criterion units which are definitely low or high. Whenever there was obvious disagreement between two or more equally qualified raters as to whether a unit was high or low, the unit was not included in the final criterion groups. Whenever there was inconsistency between a unit's average score on the ratings and on the rankings, the unit was not included in the final groups.

The field utilization of the *Safety Factors Rating* and the *Criterion Ranking* forms, by which the selection was made, had certain shortcomings. It was impossible to have each rater rate all the units—which would have been the most desirable procedure—because no raters were sufficiently familiar with the units to do so. It was a fortunate but infrequent instance when three raters could rate the same unit. For this reason, many units which, on the basis of rating and ranking scores, would appear to be definitely "low" or "high" were rated by only one person, and so had to be discarded in favor of other units where two or more raters had agreed on the unit's relative position in the installation.

The research team discovered early in the field work that some criterion raters considered themselves to be more or less qualified than other raters. Therefore, the qualifications of the raters, as they were informally expressed during the criterion sessions, were also taken into account in selecting units, that is, when there was a question about wide deviations in the scores of the units. A further consideration in the selection of units for intensive study of administrative practices (not reported in this paper) was their representativeness in terms of number and kinds of vehicles, missions of the units, and special functions or hazards.

Table 1

Means and Standard Deviations of Scores for the High, Low, and Total Groups for Criterion Rankings and Safety Factors Ratings

Groups	Rankings*		Ratings	
	M	$\sigma$	M	$\sigma$
High (16 units)	37.9	5.1	10.1	2.3
Low (16 units)	23.1	5.2	2.5	5.2
Total (93 units)	30.5	13.9	6.5	4.5

\* Each unit's rankings (by one or more persons) converted to a standard score scale that has a mean of 30 and a  $\sigma$  of 10.

The means and standard deviations of the scores of the finally selected criterion groups are shown in Table 1.

The mean rating and ranking scores of the 93 units and related data are given in Table 2.

In the selection of units, where a choice was possible, varied units were selected. This was done so as to include in the sample as many differently structured units with different missions as possible. Over-all, however, the selected high units and low units were similar.

For the high group, 9 were post units and 7 divisional units. For the low group, 10 were post units and 6 divisional units. The breakdown (Table 3) shows the make-up in more detail.

It seems to be established, therefore, that the selected high and low groups, while quite similar in types and numbers of vehicles, numbers of drivers, sizes of units, and missions, were in reality different, presumably in terms of performance.

Since the *Safety Factors Rating* was used by various personnel at the criterion sessions, it was desirable to see if there were any significant differences in the way in which various groups rated. In other words, were the ratings as a whole homogeneous? This question resolved itself into the testing of the hypothesis that the ratings of four groups of raters were random selections from the same universe. The four groups in question were: Group A. Provost Marshals, Safety Officers, and Directors; Group B. Post Ordnance, Maintenance, Transportation and Motor Of-



Table 2

Mean Ratings of Units by Higher Echelon Criterion Raters

Mean				Mean			
Unit	Rating	Rank**	Selected	Unit	Rating	Rank	Selected
1*	-2	33		48	7	22	
2*	-1	27		49	11	40	
3*	-1	18		50	2	21	
4	13	38		51	11	36	
5	7	30		52	10	33	H
6	11	31		53	10	43	H
7	10	39		54	8	25	
8	3	24		55	8	24	
9	11	40	H	56	8	24	
10	10	33	H	57	4	28	
11	7	25		58	9	34	
12	7	25		59	10	39	H
13*	5	40		60	4	25	
14*	1	43		61	9	38	
15*	3	35		62	2	18	L
16	3	20	L	63	8	31	
17*	2	25		64	8	31	
18*	4	30		65	6	35	
19	9	35		66	6	24	
20	9	25		67	5	26	L
21	14	38		68	6	36	
22	12	19	L	69	10	47	H
23*	10	25		70	11	47	H
24	-7	18	L	71	2	27	
25	9	39	H	72	6	33	
26	9	30		73	7	33	
27	10	41	H	74	-6	24	L
28	11	32		75	5	30	
29	4	16	L	76	1	26	
30	7	22		77	2	26	
31	2	25		78	7	32	
32	8	31		79	5	38	H
33	8	34		80	6	33	H
34	11	36		81	9	36	H
35	8	24	L	82	-4	14	L
36	9	30		83	4	26	L
37	None	—	L	84	8	27	L
38	13	34		85	12	33	H
39	0	19		86	10	36	H
40	5	29		87	13	28	H
41	3	24	L	88	12	33	
42	5	36	L	89	9	29	
43	3	29	L	90	12	41	
44	9	27		91	15	41	H
45	2	25		92	4	29	
46	9	33		93	4	22	L
47	4	39					

\* Only one rater rated the unit.

\*\* Converted to standard scores as in Table 1.

Table 3  
Kinds of Units in the Sample

Kind of Motor Unit	High Group	Low Group
Car Company	2	—
Truck Company	1	3
Ordnance Company	3	—
MP Units	3	1
Engineer Unit	1	2
QM Company	1	2
Signal Company	1	—
FA Battalion	1	1
HQ Company	1	1
Administrative Motor Pool	2	2
Ambulance Company	—	2
Heavy Tank Battalion	—	1
Antiaircraft Battalion	—	1

icers; Group C. Staff Officers and others not classed in Group A, B, or D; and Group D. Unit Motor Officers and NCO's.

Because of the possibility that there might be a difference in the homogeneity of ratings among the groups when rating high units as opposed to when rating low units, it was decided to analyze the ratings of high units separately from the ratings of low units. Analysis of variance was employed to test the above hypothesis. Of 16 items tested, F ratios for low units were significant at  $P < .01$  for ten items; for high units, only two items were rated differently by the 4 groups.

A second item analysis of the *Safety Factors Rating* was made in which the responses of the motor officers and motor sergeants (Group D) were compared with all of the criterion raters' responses together (Groups A, B, and C). The results of this analysis, using Strong's method (3) to obtain response weights, and testing for significance with chi square, showed that a significant difference existed between the higher echelon officers and the motor unit leaders on each of the 16 items.

It seems, therefore, that the motor officers and sergeants do disagree with the criterion raters in rating their own units. Linked with the results of the item analysis of variance, this means that motor officers and sergeants rate their units differently than do the higher

echelon officers (criterion raters). This was noticed on inspection of the response frequencies of the two groups: the criterion ratings are consistently lower than the ratings by the motor unit's own personnel. It is likely that this difference is due to a typical overrating of one's own organization which might be expected from the motor officers and motor sergeants. They are unable to place realistically their own unit in the context of units at the installation. This difference in groups, however, is much more evident in the ratings of low units than in the ratings of high units, since the analysis of variance showed 10 of the 16 items to be rated differently by the groups in rating low units, opposed to only two items when rating high units.

Leaders of low units, both officers and NCO's, are less able than high unit leaders to place their unit realistically relative to the other motor units on the post with regard to the over-all safety of the unit.

In order to determine which of the factors rated on the *Safety Factors Rating* seemed to be differentiating between high and low units, a further item analysis was made, in which the responses of the motor officers and motor sergeants from the high units were compared with those from the low units. For this comparison to be made, it was necessary to combine the ratings of NCO's with those of the motor officers. This was possible since the functions of the two groups were closely intertwined in the administration of motor pools. The item analysis comparing the responses of all motor sergeants with all motor officers (rating their own units) showed that statistically there was no reason to suppose that the former group had rated their units differently than the latter group.

The item analysis of motor officers' and motor sergeants' responses comparing high and low groups showed no marked differences in answering the questions. Essentially, this means that the motor officers and motor sergeants rate their own units the same regardless of their unit's relative position in over-all safety as rated by the higher echelon officers. This corroborates the analysis of variance results reported above, in which it was seen that



Table 4  
Percentage of Vehicles Damaged per Vehicle Operated in Period

Vehicle Type	High					Low					CR
	Number Operators	Number Damages	Per Cent Damages	Units Reporting Damages	Units Operating Vehicles	Number Operators	Number Damages	Per Cent Damages	Units Reporting Damages	Units Operating Vehicles	
$\frac{1}{2}$ ton	419	26	6.2	6	14	337	25	7.4	10	15	.65
$\frac{3}{4}$ ton	61	7	11.5	2	13	71	6	8.5	3	13	.58
$1\frac{1}{2}$ ton	73	0	0	0	7	70	20	28.6	1	7	5.29
$2\frac{1}{2}$ ton	297	15	5.1	7	15	312	19	6.1	10	16	.56
Sedans	101	14	13.9	3	4	50	8	16.0	2	3	.34
Misc.	78	9	11.5	3	12	128	28	21.9	5	13	2.02
Total	1,029	71	6.9	11	16	968	106	11.0	14	16	3.16

the unit leaders consistently rate their own units high on the *Safety Factors Rating*, regardless of where the criterion raters place the unit.

The *Vehicle Damage Report* was analyzed to determine if the numbers and kinds of damages, as recalled and reported in an interview situation by the units' mechanics, would show differences between high and low units.

The numbers of damages incurred by the high and low groups were summed separately by types of vehicles. The numbers of damages by type for each criterion group were equated to the number of vehicles of that type operated and maintained. In addition, the numbers of damages were adjusted to the

number of trips made in one month. The results are summarized in Tables 4 and 5.

Table 4 shows a significant difference between the percentages of  $1\frac{1}{2}$  ton vehicles damaged in the high group: none, as compared to 28.6% damaged in the low criterion group. It should be noted, however, that only one unit (which reported 20 accidents to vehicles of this type) caused this significant difference. The difference is seen also when relative amount of use is considered by adjusting damages to the number of trips made during the period by  $1\frac{1}{2}$  ton vehicles (Table 5).

A second significant difference occurs between the percentages of vehicles damaged in the miscellaneous category (heavy engineer-

Table 5  
Percentage of Vehicles Damaged per Trip Made during the Preceding Month

Vehicle Type	High					Low					CR
	Number Trips	Number Damages	Per Cent Damages	Units Reporting Damages	Units Operating Vehicles	Number Trips	Number Damages	Per Cent Damages	Units Reporting Damages	Units Operating Vehicles	
$\frac{1}{2}$ ton	6,007	26	.4	6	14	3,579	25	.7	10	15	1.69
$\frac{3}{4}$ ton	903	7	.8	2	13	572	6	1.5	3	13	1.03
$1\frac{1}{2}$ ton	836	0	0	0	7	980	20	2.0	1	7	4.53
$2\frac{1}{2}$ ton	3,916	15	.4	7	15	2,536	19	.8	10	16	1.85
Sedans	1,382	14	1.0	3	4	886	8	.9	2	3	.26
Misc.	893	9	1.0	3	12	2,252	28	1.2	5	13	.56
Total	13,937	71	.5	11	16	10,805	106	1.0	14	16	4.27

ing equipment, ambulances, wreckers, trailers, etc.). When relative use is considered, however, the significance of this difference disappears, since the low groups use these vehicles more frequently than do the high groups.

When the difference between percentages of all types of vehicles operated during the period is considered (Table 4), it is found to be significant. Moreover, this difference remains significant when the relative frequency of use is considered (Table 5).

These results are interpreted to mean that the low criterion units are relatively unsafe as compared to the high units. The ability of the high echelon officers to make criterion rankings and ratings in terms of safety of unit operation is substantiated and it is concluded that the subjective criterion has real validity.

#### Summary

The development of criterion measures of safety of operation for groups reported in this paper proceeded from a consideration of previous measures reported in the literature,

to utilization of rating and ranking procedures to obtain preliminary criterion groups of motor vehicle units. The criterion was not accepted as valid, however, until an investigation of damages showed a relationship to the preliminary grouping of units. It is the authors' opinion that criteria derived from ratings or rankings should be verified by showing them to be related to some critical behavioral aspects of effectiveness, acceptable to the psychologist, to the raters, and to the groups being studied.

Received April 9, 1953.

#### References

1. Johnson, H. M. The detection and treatment of accident-prone drivers. *Psychol. Bull.*, 1946, 43, 489-532.
2. Lawshe, C. H., Jr. A review of the literature related to the various psychological aspects of highway safety. *Purdue University Engineering Bulletin*, 1939, 23, 2a, Lafayette, Ind., Engineering Experiment Station, Purdue University.
3. Stead, W. H., Shartle, C. L., and Associates. *Occupational counseling techniques*. New York: The American Book Co., 1940. Appendix VI, 253-255.



## Visual Acuity Measurements by Wall Charts and Ortho-Rater Tests \*

D. A. Gordon, J. Zeidner, H. J. Zagorski, and J. E. Uhlaner

*Personnel Research Branch, TAGO, Dept. Army, Washington, D. C.*

Recently several instruments involving optical simulation of distance have been developed for large scale acuity testing. Among such devices are the Bausch and Lomb Ortho-Rater, the Keystone Telebinocular, and the American Optical Company Sight-Screener. These instruments provide means of presenting tests of right eye, left eye, and binocular acuity, as well as vertical and horizontal phoria, stereopsis, and color vision. Both near and far simulated distances may be used.

For the measurement of far visual acuity, optical instruments have several advantages over the usual method of wall chart or alley testing. The light source is "built in" and, therefore, can be made relatively accurate. Alley charts, on the contrary, vary widely in conditions of illumination. The viewing distance of instruments is achieved optically, with consequent economy of testing space. Targets may be conveniently changed without crossing the testing room. And of course, a variety of visual functions may be tested on the same instrument.

Before any one of these new instruments can be considered seriously for extensive visual testing, it should be compared with wall chart presentation. This study should be made on the basis of relative difficulty, reliability, and similarity of functions measured. The present paper deals with these problems; a comparison is made between acuity scores on wall charts and on the Bausch and Lomb instrument test.

### Review of Literature

The reliability of wall chart tests of far visual acuity has been determined (2). Data are also available on the reliability of instrument tests (1, 3). A rigorous comparison between these reliabilities cannot be made because of differences in the populations, test targets, and light levels employed. A study by Sulzman, Cook and

Bartlett (6) did employ the same subjects in comparing the reliabilities of instrument and of wall chart tests. The instruments they employed included the Sight-Screener, the Ortho-Rater, and the Telebinocular. It was found that the reliabilities of the letter wall chart tests were about the same as those of the instrument tests. They ranged from .80 to .88 for the two wall chart tests, and between .81 and .85 for the three instrument tests. In near visual acuity testing, reliabilities were also similar. The wall charts, however, seemed to be testing a visual function somewhat different from that of the instrument tests. The correlation between the letter wall chart tests was considerably higher than that between wall and instrument tests. If those correlations had been corrected for attenuation, the difference would be even larger. The authors conclude that these results may be due to the introduction of some new factor related to the optical system of the instrument or to the fact that different targets are used in the various tests.

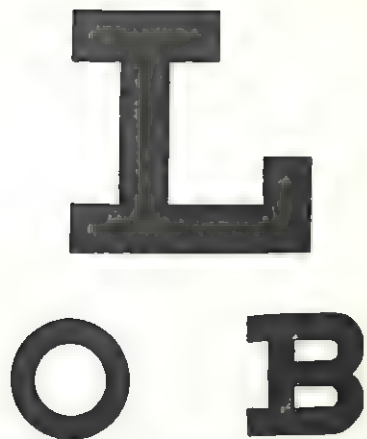
Altman and Rowland (3) determined the relationship between scores obtained on an Ortho-Rater and a wall chart when the same target was used. The wall chart was an accurate enlargement of the plate reproduced for presentation at 20 feet. One hundred and fifty-seven eyes were tested without refractive corrections in order to secure a wide range of acuity scores. A correlation of .94 was obtained between acuity scores on the Ortho-Rater and wall chart tests. This study presents supporting evidence of the identity of the visual abilities measured by the two methods.

In the present experiment, an attempt was made to compare the test-retest reliabilities and to obtain a measure of the correspondence between scores on Ortho-Rater and wall chart tests. The same subjects, targets, and light level were employed in both methods of presentation. The conditions of luminance and contrast between object and background were equalized as closely as possible. With control of these conditions, more definitive conclusions may perhaps be reached concerning the reliabilities of the two presentation methods, and the presence or absence of the "apparatus accommodation" factor thought by some to affect machine scores (5).

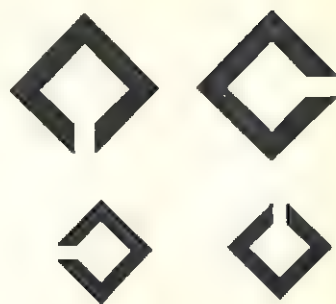
### Method and Procedure

The present experiment was conducted at the Personnel Research Branch's Pentagon Laboratory in Washington, D. C.

\* Any opinions expressed herein are those of the authors and do not necessarily reflect those of the Department of the Army.



Army Snellen



Modified Landolt Ring

FIG. 1. Visual acuity items.

The subjects were 117 soldiers from Fort Myer, Virginia. Soldiers varied in age between 19 and 37 years, with the mean age at 22.4 years, and a standard deviation of 2.6 years. The test targets were observed binocularly. All subjects who customarily wore corrective lenses used them in the experiment.

Twenty examinees who either reported having trouble with their eyes such as irritation, watering, or fatigue, or who reported having driven the night before, were considered as a special group. The decision was made to include the results of this group in the analysis after no significant differences in estimates of reliability were found between this group and the soldiers reporting no eye trouble.

The test designs included a letter chart and a modified Landolt ring chart.<sup>1</sup> Samples of the items in these targets are shown in Figure 1.

The letter chart was a modification of the Snellen chart employed by the Army in routine visual acuity examinations. Items were added to give more adequate discrimination where too few items and sizes were found on the old test. The chart consisted of 12 lines of letters ranging in size from 20/100 to 20/7.1 Snellen. The modified Landolt ring chart presented a square target rather than the circular design used in the original ring. The chart contained 11 lines of items, ranging in size from 20/135.2 to 20/5.9 Snellen.

The Ortho-Rater plates were made from the wall charts by a double reduction photographic process. It was intended to reduce the wall charts (constructed for testing at 20 feet) to .0555 of the original size.<sup>2</sup> Actually the reduc-

tion ratios, as determined by an optical comparator with microscopic attachment, are: letter plate, right eye .0546, left eye .0545; Landolt plate, right eye .0552, left eye .0549. The visual angles corresponding to the reduction ratios of the Ortho-Rater letter targets are slightly smaller than those of the counterpart wall chart; the visual angles of the Ortho-Rater Landolt targets are almost identical to their wall chart.

The laboratory in which testing took place was constructed in conformity to specifications formulated by the Armed Forces—NRC Vision Committee. The viewing distance was 20 feet for wall chart testing. Illumination was furnished by three overhead lights in flashed opal glass fixtures. These fixtures were evenly spaced along the testing alley. The front of the alley, sides, top, and floor were covered by white osnaburg cloth which served to provide an evenly lit surround over the visual field.

The brightness of the wall charts and Ortho-Rater plates was 13.5 millilamberts. A MacBeth Illuminometer was employed in making light measurements. In calibrating the brightness of the Ortho-Rater plates, observations were made against a blank plate with the eyepiece of the instrument removed. A correction was added to adjust for loss of light to be expected in transmission through the eyepiece. The required Ortho-Rater and wall chart brightnesses were secured before each session, by use of a voltmeter and a continuously variable resistance (variac).

Before being tested, each subject was shown sample targets of the designs to be used. The testing procedure was carefully explained. It was emphasized that he was to keep reading each test until told to stop. The subject was encouraged to guess if he was not sure.

The examiner observed the subject at all times to make sure that he did not squint or view the charts obliquely. The subject was rested from time to time. Responses were transmitted electrically to an adjacent room where they were

<sup>1</sup> The authors wish to acknowledge their indebtedness to Mr. Owen Conger, Typo and Design Unit, Army Publication Service Branch, TAGO, for his careful drafting of these vision targets.

<sup>2</sup> This reduction ratio is employed by the Bausch and Lomb Company in the manufacture of Ortho-Rater plates. It is based on an estimated distance of 40 mm. from lens surface to the eye, and 362 mm. from the far plate to the eye.



checked by a technician and recorded on prepared answer forms.

The following presentation order of tests was maintained: wall chart letter, wall chart modified Landolt, Ortho-Rater letter, Ortho-Rater modified Landolt. These tests were a portion of a larger group of 17 mesopic and photopic targets given in the same session. Subjects had observed five mesopic wall charts and two mesopic Ortho-Rater plates before taking the four tests discussed here. The letter wall chart was the third test given on the photopic level, the modified Landolt wall chart was the fifth, the Ortho-Rater letter plate was the eighth, and the Ortho-Rater modified Landolt plate was the ninth of the ten tests given at the photopic level. The same procedure was followed in the retest session two weeks later.

Results

An indication of the relative difficulty of wall chart and Ortho-Rater presentation is shown in Table 1. The mean represents the average number of items achieved by the subjects before the criterion of failure was met. These results are presented for four scoring methods:

(a) Number of rights before two consecutive miscallings were first made; (b) Number of items attempted before two consecutive miscallings were first made; (c) Number of

rights before three consecutive miscallings were first made; and (d) Number of items attempted before three consecutive miscallings were first made. These were utilized to show the effect of scoring method on results and, thus, give the results wider generality. It will be recognized that these scoring methods are non-independent measures.

It may be seen that subjects were able to read further on the wall chart letter tests than on the Ortho-Rater letter plates before meeting the criterion of failure. This difference in difficulty may perhaps be explained by the somewhat larger visual angle of the letter wall charts (see Method and Procedure). The scores on the Landolt tests, where more perfect reproduction of visual angle was achieved, are about equal for the two methods of presentation. The standard deviations are approximately the same, except that the Ortho-Rater Landolt retest shows greater variability than its wall chart. Although there are several significant differences in means and standard deviations of the two methods of presentation, the differences are too small to be of practical importance. In Snellen acuity units, negligible changes in scores are implied. As shown in

Table 1  
Comparison of Means and Standard Deviations for Wall Chart and Ortho-Rater Tests (N = 117)

Target	Scoring Method	Mean		t* Ratio	Standard Deviation		t* Ratio
		Wall Chart	Ortho-Rater		Wall Chart	Ortho-Rater	
Letter (Test)	A	62.6	62.0	1.00	11.0	11.1	0.15
	B	64.8	64.2	1.14	11.0	11.5	0.82
	C	65.0	63.3	3.07	10.4	11.1	1.46
	D	68.9	66.3	4.53	10.8	11.8	1.74
Letter (Retest)	A	63.3	61.4	3.61	11.5	11.6	0.32
	B	65.6	63.7	3.26	12.2	12.0	0.40
	C	65.7	63.6	3.94	11.0	10.8	0.32
	D	69.6	67.6	3.84	11.5	11.2	0.63
Landolt (Test)	A	60.9	61.0	0.19	11.7	12.5	1.10
	B	62.3	63.0	0.90	12.0	12.8	1.07
	C	63.3	63.5	0.26	11.3	12.2	1.35
	D	66.8	67.5	0.94	12.0	12.9	1.31
Landolt (Retest)	A	62.0	62.7	1.12	11.3	13.2	2.96
	B	63.8	64.7	1.44	11.7	13.2	2.31
	C	64.3	65.1	1.15	11.3	13.7	3.64
	D	67.8	69.1	1.71	12.3	14.6	3.32

\* A t ratio of 1.96 indicates that the difference obtained is significant at the 5 per cent level of confidence. A t ratio of 2.58 indicates that the difference obtained is significant at the 1 per cent level of confidence.

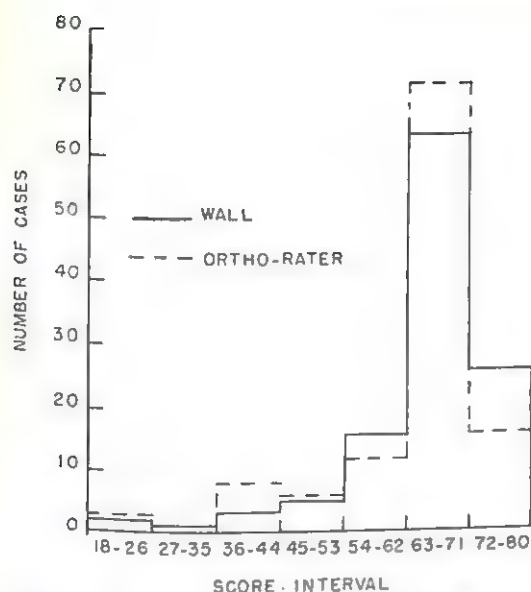


FIG. 2. Distribution of scores on the new Army Snellen tests. Items 30-39 of the tests are 20/20 acuity value ( $N=117$ ).

Figures 2 and 3, the test distributions are very similar. In general, the evidence does not indicate that Ortho-Rater and wall chart presentations differ greatly in difficulty and variability.

The test-retest reliabilities of wall chart and Ortho-Rater scores are shown in Table 2. All Ortho-Rater reliabilities, with one exception, are significantly higher than those of the wall charts.

The higher reliabilities of the Ortho-Rater plates cannot be explained by the fact that the Ortho-Rater plates were administered after the wall charts. If increased reliability is associated with later tests administered in

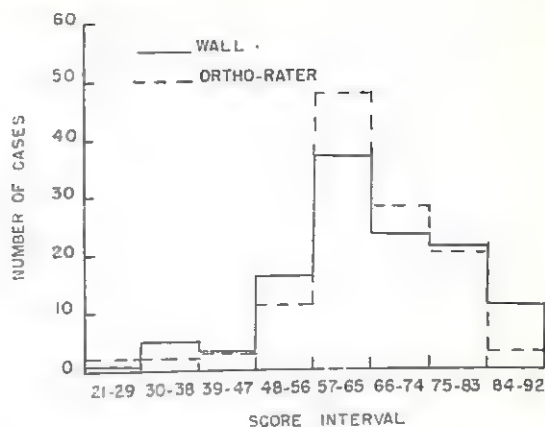


FIG. 3. Distribution of scores on the modified Landolt ring tests. Items 38-45 of the tests are 20/20 acuity value ( $N=117$ ).

the session, the Armed Forces Far Visual Acuity test, administered twice, should have shown this effect. This test was administered as the second and tenth (last) test of the photopic series. The test administered in the last position shows a significant decrease in reliability for two scoring methods and a non-significant increase for two other methods.

The correlations between scores on the wall charts and on the Ortho-Rater plates are presented in Table 3. These correlations are based on scoring method C, which was the most reliable method employed (see Table 2). They are about as high as the test-retest reliabilities. The mean of the correlations is equal to .83; the mean of the reliabilities of scoring method C is equal to .85. The mean of the correlations, corrected for the attenuating effects of unreliability in each variable, is

Table 2  
Wall Chart and Ortho-Rater Test-Retest Reliabilities ( $N=117$ )

Scoring Method	Letter Test		$t^*$ Ratio	Landolt Test		$t^*$ Ratio	Far Visual Acuity Test		$t^*$ Ratio
	Wall	Ortho-Rater		Wall	Ortho-Rater		Second	Tenth	
A	.81	.90	3.30	.73	.81	1.94	.90	.83	2.72
B	.78	.89	3.69	.69	.79	2.14	.88	.81	2.43
C	.88	.92	2.04	.75	.85	2.82	.81	.86	1.55
D	.80	.87	2.23	.65	.79	2.91	.80	.81	0.29

\* In this context a  $t$  ratio of 2.58 indicates that the difference is significant at the 1 per cent level of confidence. A  $t$  ratio of 1.96 indicates that the difference is significant at the 5 per cent level of confidence. In computing the  $t$  ratios, the correlation between  $r$ -transformations of the reliabilities was estimated at .40 for all comparisons by an approximation formula described by McNemar (4, p. 125). A standard error of .103 was obtained from the formula  $(\sigma_{z_1 - z_2} = \sqrt{\frac{2 - 2r_{12}}{N - 3}})$  used in obtaining the  $t$  ratios.



Table 3  
Correlations between Wall Chart and Ortho-Rater Tests (Scoring Method C)<sup>1</sup> (N = 117)

Variables	Test Session	Retest Session	O-R Test Session and Wall Retest Session	Wall Test Session and O-R Retest Session
Wall chart vs. Ortho-Rater (Letter)	.85 (.94)	.87 (.97)	.86 (.96)	.89 (.99)
Wall chart vs. Ortho-Rater (Landolt)	.78 (.98)	.84 (1.00)	.77 (.96)	.80 (1.00)

<sup>1</sup> Correlations corrected for attenuation are given in parentheses.

.98. These data offer little evidence to support the existence of a machine factor or "apparatus accommodation factor" specific to Ortho-Rater presentation.

### Discussion

The finding that Ortho-Rater tests are more reliable than wall chart tests presents a problem for interpretation. The superiority of instrument presentation may be due to lowered visual distraction with limitation of the surround, or to some other advantage of subject or stimulus characteristic leading to greater constancy of conditions. It is believed that the difference in reliability between Ortho-Rater and wall chart presentation will be even greater in operational testing than that found here. It is well known that the conditions of wall chart testing differ widely from place to place.

If visual angle, background luminance, and contrast between object and background are equated as closely as possible between Ortho-Rater and wall chart presentations, closely equivalent measures are obtained. The difficulties of the tests and the variability of scores are similar. When the correlations of the tests are taken into consideration, the methods appear to measure the same visual abilities.

### Summary

This study presents a comparison of visual acuity scores obtained on Ortho-Rater plates with visual acuity scores on duplicate wall chart tests. A total of 117 subjects were tested binocularly and retested two weeks later. Letter and modified Landolt ring tar-

gets were employed. Previous practice had been given on other mesopic and photopic wall chart and Ortho-Rater plates before the tests under consideration were given.

The following results were obtained:

1. The two methods of presentation were of equal difficulty, except for slight discrepancies introduced by photographic reduction.
2. The reliabilities of the Ortho-Rater tests were significantly higher than those of the wall chart tests.
3. The correlations between Ortho-Rater and wall chart tests were about as high as the reliabilities of the tests themselves. When corrected for attenuation, these correlations approach unity. No evidence is afforded, under these conditions, of a machine or "apparatus accommodation" factor affecting Ortho-Rater acuity scores.

Received April 11, 1953.

### References

1. Adams, J. K., Beier, D. C., and Imus, H. A. *A test-retest reliability study of the Bausch and Lomb Ortho-Rater with naval personnel*. OSRD Report No. 3969, Aug. 1, 1944, Applied Psychology Panel, NDRC, 1-32.
2. Adjutant General's Office. *Studies in visual acuity*. Wash., D. C.: U. S. Gov't. Printing Office, 1948.
3. Altman, A. and Rowland, W. M. Measures of acuity with optical simulation of distance. *Quart. Rev. Ophthalmol.*, 1952, Vol. 8, No. 1.
4. McNemar, Q. *Psychological statistics*. New York: Wiley, 1949.
5. Sloan, L. L. Measurement of visual acuity. *Arch. Ophthalmol.*, 1951, 45, 704-725.
6. Sulzman, J. H., Cook, E. B., and Bartlett, N. R. The reliability of visual acuity scores yielded by three commercial devices. *J. appl. Psychol.*, 1947, 31, 236-240.

# Effect of Illumination on Scores with Instrument Acuity Tests

Newell C. Kephart and Stanley Deutsch

*Occupational Research Center, Purdue University*

Standardized tests for the measurement of visual acuity have often been administered under conditions which were not held constant, both within tests and between tests. The standardization procedure has been violated in many different ways; varying light conditions, the uncontrolled mixture of artificial and sunlight, angles of view, and the like. Self-contained instruments such as the Bausch and Lomb Ortho-Rater help to overcome these undesirable deviations in practice which serve to reduce the validity of the measurements (1, 2).

However, it is realized that despite precautions, occasional variations in the source of power in the industrial establishment may lead to greater or lesser illumination even within the Ortho-Rater than that considered to be standard for testing visual skills. This experiment was designed to determine whether or not such deviations of lighting present serious drawbacks to obtaining accurate measures of visual acuity.

## Procedure

A total of 55 college students in a course in general psychology were tested on a standard Ortho-Rater, using standard practice for administration (3). The illumination in the instrument was varied by means of a Variac rheostat. The only deviations from standard testing procedure were these changes in the illumination levels which were artificially induced external to the instrument to simulate conditions which might occur in an industrial situation.

Acuity was measured at both the near and far point and a total of five levels of illumination was used at each distance. Since the Ortho-Rater had just been checked and the bulbs replaced where necessary, the normal level of illumination was used as a base of 100. Using an illumination level meter to determine amount of illumination, three lesser amounts of illumination and one greater quan-

tity were obtained by means of the rheostat. This procedure provided illumination for the targets in the following percentages of standard for far acuity: 12, 56, 75, 100, and 125. For near distance the percentages were: 10, 46, 75, 100, and 125 of the normal illumination.

The stimuli used were the "both eyes" and the "right eye" targets. The left eye was occluded at all times by a mechanical device built into the machine. The targets and levels of illumination were presented in a random order throughout the experiment, and were changed for each subject. Although only the right eye saw the material the "both eyes" targets were used in addition to the "right eye" targets in order to provide additional data.

## Results

The mean acuity scores obtained by the 55 subjects are shown in Table 1. For the target at the optical distance of 26 feet, statistically significant differences at the 1% level of confidence were obtained only for the targets employing 10 and 12% of the standard illumination. For all of the presentations where the illumination was 46% or greater, no statis-

Table 1

"p" Ratios Between Acuity Test Scores at Various Levels of Illumination and a Standard Level of Illumination

Level of Illumination	Target					
	Both Eyes			Right Eye		
	Far	Near	Level	Far	Near	Level
10%		11.21	1%		11.27	1%
12%	5.91		1%	3.37		1%
46%		7.38	1%		3.89	1%
56%	1.00		N.S.	.65		N.S.
75%		1.35	N.S.		1.18	N.S.
75%	.54		N.S.	.61		N.S.
125%		.74	N.S.		1.56	N.S.
125%	1.56		N.S.	1.29		N.S.



tically significant differences were found for this group.

The decrease in illumination for the near targets (distance of 13 inches) was somewhat more critical. When the lighting was reduced as low as 46% or less of standard, differences were obtained at the 1% level of confidence. No significant differences were found for those targets receiving 75% or more illumination.

When 125% of the prescribed lighting was employed, the differences from normal illumination were not significant.

#### Discussion

The study was undertaken to determine whether significantly different results would be obtained if the levels of illumination were increased or decreased from that established by the manufacturers of the Ortho-Rater. Several factors which might potentially invalidate the results of instrument acuity tests can be hypothesized. These might consist of temporary fluctuations in the power supply in a factory test room, reduced efficiency of the Ortho-Rater light sources prior to examination, or any feature which might change the illumination level within the instrument.

Effects of such deviations, however, appear to be minimal. A decrease of more than one-fourth is required before the differences in results become meaningful.

It is of interest to note that near acuity suffers more readily than far acuity. Whether this indicates that near visual tasks require more illumination than far visual tasks cannot be answered from the information presented here.

The 25% increment in illumination used to augment the standard at no time produced a significant change in visual acuity scores.

It is a standard procedure for Ortho-Rater operators to check the operational condition of their instrument prior to use. As a consequence, chances of using a bulb operating at only 50% of efficiency are slight. Deviations in the power supply of this magnitude are instantly obvious, and testing should await the return of the more nearly normal level of illumination. This study demonstrates that minor changes in illumination do not appear to have any real effect upon the test results.

#### Summary

1. Decreases in illumination as great as one-fourth of standard did not affect scores on visual acuity with the Ortho-Rater.

2. Increases in illumination as great as one-fourth of standard did not affect these acuity scores.

3. Near acuity scores suffer to a greater degree than far acuity scores when illumination is decreased more than 25 per cent.

Received March 18, 1953.

#### References

1. Feinberg, R., and Wirt, S. E. Visual acuity in relation to illumination in the Ortho-Rater. *J. appl. Psychol.*, 1947, 31, 406-412.
2. Jobe, F. W. Instrumentation for the Bausch and Lomb Industrial Vision Service. *Bausch and Lomb Magazine*, 1944, 20, 6-7, 14-15.
3. *Standard Practice in the Administration of the Bausch and Lomb Occupational Vision Tests with Ortho-Rater*. Bausch and Lomb Optical Co., 1944.

## Applied Psychology in Action

### Psychological Research in Personnel Administration

Joseph G. Colmen

*Civilian Personnel Research Branch, Headquarters, USAF*

Industrial management has undoubtedly been skeptical about the value of the personnel psychologist as a direct part of its operations payrolled as its job evaluation, training, organization and methods and other functions are. Yet a large number and variety of management problems can be attacked by the application of the specialized skills of the research psychologist. And in most cases not only can they provide the most valid solutions and recommendations but can do this in a manner which will please even the most practical administrator. To do this, it seems important for the research psychologist to be close enough to the management and operations of the organization so that he can sense needs for research in day-to-day problems. And he can make acceptable recommendations for application of research results in the same setting. The possibilities for success are greater, of course, where the relationship between administrator and psychologist is a close and continuing one.

The Civilian Personnel Research Branch (CPRB) of the U. S. Air Force Headquarters is in the fortunate position of approximating this ideal. This Branch conducts psychological research originating from everyday problems of the civilian personnel program of the Air Force.

Many problems had pointed to the need for test development when the CPRB was established in 1950. One of the urgent problems was to find more objective means for determining supervisory potential to improve the level of supervisory proficiency within the Air Force, especially in the important area of human relations. Another problem was to find an objective measure of potentiality for administrative work as a basis for selecting junior employees for specialized development and training.

These problems demanded pioneering research in areas where success had been previ-

ously only hopefully promising. Fortunately, management recognized that scientific research took time. It did not insist on "quick" results at the sacrifice of adequate research.

As work on these basic problems progressed, other needs became evident. It was noted that from time to time Air Force installations were conducting attitude surveys among civilians without the benefit of instruction in accepted principles of public opinion polling. Guidance was needed in formulating questions, sampling employees, conducting surveys, and analyzing and interpreting results. A compact, highly functional guide to the conduct of civilian employee attitude surveys was prepared to correct these deficiencies together with a questionnaire with general applicability at all Air Force bases.

Because test administration would be undertaken by all Air Force bases upon completion of the research, a guide on establishing test administration facilities and on the best methods of administering and scoring tests was developed. Also, regulatory material was published to assure effective coordination and highest quality of research throughout the civilian personnel program and to stimulate necessary personnel research where resources permitted.

To conserve research resources, use was made wherever possible of the work of other organizations with adaptation and special check and validation being accomplished within the Air Force. On the other hand, no test is authorized for use without specific validation on the group for which it is intended or for a purpose other than that for which it was validated.

With what has been a very heavy schedule, new developments in technical areas have still been possible. Readers of this paper trained in test theory will recognize that depth, quality, and originality of research have not been sacrificed even in the applied setting in



which it is conducted. A nomograph for determining significance of difference between percentages for finite populations was developed for handy use by statistically untrained people in analyzing attitude survey data. An "unconventional" key and rationale for it was hypothesized and verified as a valid keying technique for personality test items. The methods of weighting tests by precise Wherry-Doolittle beta weight methods or Wherry-Gaylord integral weighting were found to be less appropriate under certain circumstances than mere unit weighting of tests so selected. Later discussions with Dr. Wherry have confirmed that the number of cases and size of test intercorrelations do affect the stability of weights derived by these methods.

Awareness of the specialized skills brought to the organization by research psychologists soon led personnel specialists in other program areas to seek the services of the Civilian Personnel Research Branch. Typical assignments in other areas of applied research were: to determine the advantages and limitations of functional music in the work setting; to develop sampling methods in connection with interviews used as part of an evaluation of the effectiveness of civilian personnel programs at Air Force bases; to determine the best sources and methods of ascertaining supervisory training needs as a basis for developing training course content; to determine the extent and causes of turnover among civilian personnel office professional staffs; to develop a battery for selection of fiscal accounting clerks; to evaluate effectiveness of employee suggestion systems; to make recommendations concerning the most valid and reliable information about personal characteristics of candidates for positions from inter-

views, vouchers and other screening methods; with the Civil Service Commission and four other federal agencies, to develop selection methods for reducing the number of emotional misfits finding their way into overseas jobs; and others.

Though research has constituted the major responsibility of the CPRB, it has never divorced itself from the personnel administration of which it is a part, so that research needs are perceived by the psychologist in the operating and staff civilian personnel problems with which he is in close contact. Nor is the work completed when research findings are reported. Instead, implementation of those findings in the practical setting of the operating civilian personnel office becomes in large part also the responsibility of the researcher. And he is kept informed of and asked to comment on personnel administration programs, policies and procedures which are under consideration in the Directorate of Civilian Personnel.

The satisfaction of management with the accomplishments of its personnel research function is seen in continued support and growing acceptance. By keynoting economy and improvement of operations, data have been accumulated showing how the results of the work of CPRB have much more than offset its modest cost. It is interesting to note that whether or not a personnel research activity is maintained, management will conduct research studies. Staffing with personnel specifically trained for such work pays dividends, if in no other way than in making such research sufficiently sound to assure management that the conclusions may be applied with confidence.

## Time Limit versus Work Limit Methods of Test Administration

E. B. Knauff

*Aetna Life Affiliated Companies, Hartford, Connecticut*

The majority of mental alertness tests used in the employment situation are speed tests in which the time factor plays an important role. Those of us working in business and

industry are sometimes asked whether such speed tests unduly penalize the "slow and accurate individual" who might be a very satisfactory worker.

Some information relative to this question was obtained in the process of revising the LOMA-1 Test. This is a 15-minute general mental ability test of 236 items (including number series, same-opposites, analogies, and general information) developed by the Life Office Management Association for use in member companies. Many studies during the past 15 years have demonstrated the validity of this test as an aid to selection of clerical employees in insurance companies. A revised form of this test was recently administered to employees who took the test with the usual 15-minute time limit and then were permitted to complete the remaining items with time noted, but no time limit.

Data based on 235 employees in four dif-

ferent life insurance companies showed that scores obtained in the 15-minute time limit correlate + .88 with scores on the entire test obtained under untimed conditions. The mean time required to complete all items was 30.1 minutes.

For this sample, it appears that individuals performing relatively poorly on a mental alertness test under timed conditions will not appreciably change their standing in the group when permitted to complete the test with no time limit. The 235 employees represent a sample of persons hired within the past five years and still employed by the four companies. The great majority are between the ages of 18 and 35 and are high school graduates.

---

## Employee Opinion Surveys

"Why should we invite employees to criticize us? They do enough of that anyway without being asked."

That's the attitude of many top management officials. . . .<sup>1</sup>

But San Diego Gas and Electric Co. is one company that calculated the risk. . . . Now it says that it's glad. One reason SDG&E is happy with the results is that the employees gave the company a pretty good rating.

<sup>1</sup> See McMurry, R. N. Management's reactions to employee opinion polls. *J. appl. Psychol.*, 1946, 30, 212-219. (Reference added by Editor.)

But they received some very specific suggestions. . . . A total of 3,380 unfavorable comments and 1,290 suggestions were made by 2,178 employees. . . . The company had determined in advance to do something about reasonable complaints and it followed up quickly. Top management gave full approval and support. Adequate assurance of anonymity was provided by the Industrial Relations Section of California Institute of Technology which conducted the survey getting a 99 per cent return. (*Condensed from Business Week*, November 7, 1953, p. 167.)



## Book Reviews

Lawshe, C. H. *Psychology of industrial relations*. New York: McGraw-Hill Book Co., 1953. Pp. vii + 350. \$5.50.

During the past decade psychology has had a rapidly increasing impact upon industrial management through two developments: (1) the introduction of psychologists into staff, and occasionally line, positions; (2) the training of various levels of management in the principles of human behavior. This has led to two types of publication: those for use in training industrial psychologists and those which present the findings of psychology to the non-psychologist. This volume is an example of the latter.

Seven authors are involved—two from Purdue (Lawshe & E. J. McCormick), one each from the Army (A. J. Drucker), Air Force (W. F. Long), and Navy (E. E. Dudek), and two from industry (K. Oliver & R. I. Dawson). The fifteen chapters deal with the usual topics: principles of human behavior, motivation, attitudes, placement, training, supervision, employee complaints, counseling, efficiency, wage administration, employee and employee-management relations. Although the chapters were written by individual authors, they are remarkably similar in style, level of reading difficulty, and point of view and there appears to be relatively little overlap of content except when desirable. In general, the writing is clear and direct, with attempts to define psychological terms by means of industrial examples. Each chapter concludes with a set of references to journal articles or psychological texts.

So much for the over-all structure of the book. The question remains—how well has the goal of communicating industrial psychology to the non-psychologist been achieved? The only true answer to the question can come from actually measuring what effect the study of this book by non-psychologists has had upon their knowledge, skills, and attitudes in human relations. Lacking research findings, the reviewer can merely speculate as to the book's probable value.

The problem in writing for the non-psychologist is, of course, deciding what one wants to communicate. There are several possibilities: (1) psychological findings with

or without the underlying evidence; (2) suggestions for *what to do* in industrial settings, with or without reference to the principles being applied; (3) a *point-of-view*, derived from psychological principles of human behavior.

In this reviewer's opinion, the authors have done a creditable job in presenting a large body of facts and principles, backed up with sufficient references to research literature. However, certain areas to which psychologists have devoted considerable thinking and research are inexplicably omitted or merely mentioned in passing, viz., industrial safety, democracy in management, executive development, employee rating methods, characteristics of the learning curve, transfer of training.

It is difficult to evaluate the "how-to-do-it" aspect of this book. There appears to be an attempt to present principles, not specific applications; the discussions tend to include more of the "it's important to take the following things into account" type of statement than to describe *how* to take them into account. There is a somewhat too frequent dependence upon a brief raising of a question or listing of factors and then a reference to a bibliographic item for the details, assuming that the reader will go to the sources.

Possibly psychology really cannot give very many specific suggestions for industrial practices and that its real contribution is in methodology and point of view. If so, this book serves a useful purpose in getting across to the non-psychologist the basic attitudes towards human problems which characterize psychology, e.g., "emphasis upon the people that work rather than upon the product they make," the importance of satisfying basic human needs, the need for a "basic respect for human beings and a genuineness of purpose in dealing with employees." To the extent that publications of this type stimulate operating personnel to examine their fundamental attitudes toward human behavior they will facilitate the acceptance of programs developed by the industrial psychologist and raise the level of daily interpersonnel relations.

A. S. Thompson

Teachers College,  
Columbia University

McFarland, Ross A. *Human factors in air transportation*. New York: McGraw-Hill, 1953. Pp. xv + 830. \$13.00.

The advent of the airplane and the extension of its performance characteristics have subjected those occupying this device to an unprecedented variety of environmental stresses. Extremes and variations of temperature and accelerative forces are commonplace in military flight. The tasks of maintenance and operation have required the development of skills unthought of fifty years ago. In many ways, the airplane has provided a laboratory and a never-ending set of problems for the engineer, the physiologist, and the psychologist.

Drawing upon his unique and extensive acquaintance with practically every aspect of commercial and military aviation, Dr. McFarland has written an encyclopedic volume of over 800 two-column pages, illustrated with well-selected tabular, graphic, and pictorial displays. Each chapter is followed by a selected bibliography.

*Human Factors* covers thoroughly the areas of selection, maintenance of proficiency, and safety which are implied by its title as well as such topics as sanitation and health in airline operations, the care of passengers, and a description of medical programs. The discussion of physical factors involving circulatory and sensory phenomena is unusually comprehensive.

The reviewer is impressed by the clear style, careful organization, and excellent typography of the book. This work would seem to be not only a landmark in the area for which it is intended but also a valuable source of information for those concerned with most branches of personnel psychology. Its chief drawbacks are likely to be its size and perhaps its price of \$13.00, although the latter is certainly modest for so large a book aimed at a limited audience.

George K. Bennett

*The Psychological Corporation,*  
New York, New York

Husband, Richard W. *The psychology of successful selling*. New York: Harper and Brothers, 1953. Pp. 306. \$3.95.

This book is directed to all salesmen to aid them in their daily work. Its emphasis is

on sales tactics, from finding your prospects through approaching him and overcoming his resistance to closing the sale. There is also a short section concerning the selection of salesmen, helping him to compare his traits with those of successful salesmen.

This book is not intended to be a professional book for psychologists; rather it is deliberately designed to be easy, informal reading without technical language or reference to experiments or statistics. It is admittedly based upon reading leading books by professors of business and sales personnel, sales journals, trade publications, newspapers, popular magazines, training manuals of certain companies, and the author's personal experience. Drawing upon these sources, the book presents a series of rules, principles, steps, and laws on how to be effective in each phase of selling. These are liberally illustrated with clever examples and entertaining anecdotes. Sprinkled with this is advice and moralizing based on the personal opinion of the author.

Thus while the author claims this is the first general book on salesmanship written by a professional psychologist, it is certainly free from the concepts and language of the psychologist. You will find no discussion per se of motivation, adjustment, individual differences, learning, perception, and so forth. This was apparently deliberately omitted in order to make the book more appealing and readable for salesmen. There are many paragraphs, however, in which the oversimplification has led to statements which the reviewer could not accept. The book is a challenge to psychologists in that it reveals a large area in which practical applied research can still pioneer. Many statements of the book are based on what "should be" and reasoning from analogy; it would be quite difficult to support them with evidence from scientific references.

In general, there is little in the book to recommend it even to sales managers or salesmen over the many other volumes written in this field.

Brent Baxter

*The Prudential Ins. Co. of America,*  
Newark, New Jersey



Jahoda, Marie, Deutsch, Morton, and Cook, Stuart W. *Research methods in social relations, with especial reference to prejudice*; Vol. I: *Basic processes*; Vol. II: *Selected techniques*. New York: The Dryden Press, 1951. Pp. x + 421, x + 423-759. \$6.00 (set).

As indicated in the preface of this publication, "This book is in many ways the outcome of group effort. The idea of producing it arose in a group; it is presented under the auspices of a group; its production was financed by several groups; it had the editorial guidance of a group; and it was produced by a group." The sponsoring agency for the book was The Society for the Psychological Study of Social Issues.

The two volumes themselves show the impacts of their sponsorship, and of the many hands which have been laid upon them. There is a sense of urgency in the book's treatment of problems of social intolerance and discrimination, and an implication of mild, but persistent, exhortation to the reader to take constructive steps in combating these evils. For the scientific reader the proper course is to be found in "action research" (participative research directed toward the solution of tangible problems), and for the social practitioner the recommended course is cooperation with the scientific investigator.

In spite of the instances of apparent overearnestness and occasional naiveness which occur in the book, it still remains a useful and informative document. The sections on research planning, on practical issues in research, and the uses and applications of research results are admirable. The second volume, which consists for the most part of separate papers by various contributors, should also be noted. Readers desiring short, but critical and dependable resumés of topics such as scaling concepts, the use of panels, and sociometric analyses will find this volume a valuable reference.

The avowed purpose of the book was to reach two audiences: the conductors of social research and the users of social research. In the reviewer's opinion neither of these specific goals has been satisfactorily met, but another, equally legitimate goal has been. The book is too superficial and given over to

standard illustrations to be of much help to the practicing researcher, and seems to be too academic and technical to have much appeal to the practical man-of-affairs. But the book does have a thoroughness, and an informative and dependable quality which would make it an excellent source book for non-specialists, and for students who wish to gain a brief, but competent and comprehensive overview of this field of research.

Harrison G. Gough

University of California,  
Berkeley

Coombs, C. H. *A theory of psychological scaling*. Ann Arbor: University of Michigan Press, May, 1952. Pp. vi + 94. \$1.75.

If you've had a hard day measuring attitudes, don't expect this small monograph to provide an evening's relaxation. It's packed from cover to cover with non-superfluous material. It is to the author's credit that he has said so much in so short a space; nevertheless, persons lacking expertness in scaling theory will not digest the contents properly. On the other hand, scaling theorists will accept this tidbit as a juicy morsel and will soon be looking for more.

The theory presented here has been in the process of formulation for four years and represents the contributions and criticisms of many scholars. It has undergone continuous modification in response to these criticisms and will undoubtedly undergo more. However, its publication now "is necessary for the presentation of certain consequences of practical interest to psychologists and social scientists" (page v).

Roughly speaking, the presentation is made in four parts: (1) a general discussion of the aspects and problems of psychological measurement; (2) a listing and brief explanation of the definitions and postulates on which the theory is based; (3) the development and interpretation of genotypic and phenotypic parameters; and (4) derivations of the consequences of various genotypic conditions and the application of the theory to several sample experiments.

The effort has been to present a mathematical model which will satisfy observed behavior. As such, the theory must resolve certain fundamental issues such as the question

of defining a psychological trait in a mathematical sense. In order to resolve problems of this sort, parallel systems have been developed—one formulated at the genotypic level and referring to an individual's inferred, underlying abilities and behavior—the other formulated at the phenotypic level and referring to an individual's observed, manifest behavior. The ultimate objective of the system, then, is to treat information obtained from a set of phenotypic observations so as to allow inferences at the genotypic level.

The first five chapters present the theoretical framework for the realization of this objective. It is in Chapter VI that Dr. Coombs discusses the area of joint scales—the final relating of the phenotypic to the genotypic. And it is here that he must confess partial defeat, for he comes face to face with the problem of the *direction* of the inferences we, as theorists, wish to make. Thus, starting with certain genotypic conditions, it is demonstrated what the consequences must be in terms of manifest behavior. Unfortunately, in the practical situation, it is only a hope that we may apply these relationships in the opposite direction. From characteristics of manifest behavior, we desire to infer characteristics of behavior at the genotypic level. The author sums up the difficulty, "It has not been shown that for this given set of parameters or characteristics of the manifest data it is *necessary* that these and only these conditions must characterize the genotypic level" (page 52).

The author openly states that the theory is not in final form. By implication, it is his hope that this publication will initiate interest resulting in a wider range of development for the theory in both its abstract and real aspects. To this end, the monograph represents a good start.

Marvin D. Dunnette

*The University of Minnesota*

New York Academy of Medicine and the Josiah Macy, Jr. Foundation. (*Transactions of the Conference on) Morale—and the prevention and control of panic.* New York: New York Academy of Medicine, no date. Pp. 75. No price cited.

This publication is aimed at inspiring widespread consideration and study of morale and

panic. Its audience is not defined but it appears to be officials who may have responsibility for controlling public morale and panic.

The purpose of the conference was to explore available knowledge of the problems and discover a way of implementing the pooled knowledge through action. Conferees were 1 Ph.D., a psychologist in a Veteran's Hospital; 8 M.D.'s, psychiatrists from schools, state and national medical associations and governmental agencies; and 2 representatives of public information media, an official from a radio broadcasting company and the editor of a city newspaper.

The conferees earnestly advocate study of the problems and use of the resulting findings. Meerloo, formerly chief of the Psychological Department of the Netherlands Army, contributed most of the specific references to evidence on the factors affecting and means of controlling panic, drawn from his own work during World War II. Herbert Brucker, Editor of the *Hartford Courant*, emphasized the belief that factual, play-by-play reporting of the news events as they occur is probably the greatest contribution public news media can make. He argues against attempted manipulation of the news and exhortatory releases from government officials as having less than good effect upon public morale.

Much conference time was devoted to recital of personal experiences and to reference to incidents ranging from biblical events to postwar reactions of German war leaders. Various ways of controlling morale and panic with varying degrees of success were cited with Meerloo's practical findings being of greatest interest.

To focus the attention of public officials, educators, and research personnel on these problems appears desirable. Perhaps a joint attack, by experimental study of isolated factors and by concurrent multifactor study with techniques such as were described during the first meeting of the Operations Research Society of America, would be productive.

In summary, there was little experimental evidence or firm knowledge about causes and control of public morale and panic disclosed during the conference. It is not clear that the aim of inspiring widespread consideration



and study of morale and panic will be attained by publication of the transactions of the conference.

Clark L. Hosmer

*United States Air Force*

Traxler, Arthur E., Jacobs, Robert, Selover, Margaret, and Townsend, Agatha. *Introduction to testing and the use of test results in public schools*. New York: Harper and Brothers, 1953. Pp. 113. \$2.50.

This book is designed to serve as a "practical, down-to-earth handbook for schools beginning the use of objective tests, for teacher discussion groups, for in-service training programs, for persons who have had experience with tests but who desire to brush up on the simpler fundamentals of testing, and for introductory classes in tests and measurements." It is a revision of Educational Records Bulletin No. 55, *Introduction to Testing and the Use of Test Results* (Educational Records Bureau, 1950) which was prepared primarily for independent schools.

A general discussion of the role of objective tests is followed by sections on planning a testing program, selection, administration, and scoring of tests, and analysis, interpretation, recording, and use of test results. Elementary concepts from test theory and statistics are presented in context. Illustrative material is utilized extensively; interpretations of data for individual students and classes, and copies of score reports, cumulative record forms, and the like make up a substantial portion of the book. Throughout, the authors give detailed attention to the limitations of objective tests and to cautions which should be exercised in the interpretation of test results.

Each chapter includes a list of references for readers who wish to go beyond this introductory handbook. For such readers, supplementary information on score interpretation is likely to be of special concern; the use of test results for descriptive and comparative purposes is treated more explicitly than is their application in predicting future performance.

This brief, nontechnical book should be distinctly useful to the groups of readers toward whom it is directed. Despite its title,

the revision seems equally appropriate for public and independent schools. From the standpoint of the former, the more detailed discussions of test selection and program planning included in the revised edition should be of particular interest.

Marjorie Olsen

*Educational Testing Service,  
Princeton, New Jersey*

Powers, Edwin and Witmer, Helen. *An experiment in the prevention of delinquency: The Cambridge-Somerville youth study*. (With foreword by Gordon W. Allport.) New York: Columbia University Press, 1951. Pp. xliii + 649. \$6.00.

The book is devoted to the description and evaluation of a program, or as the authors call it, an experiment in the prevention of juvenile delinquency. The program had its origin in an idea formulated by Dr. Richard Clarke Cabot. He believed that the delinquency of boys could be prevented were it made possible for them to come under the constructive influence of friendly counselors.

The study was begun in 1935 and terminated in 1945. In design, the study as initially conceived was in the best scientific tradition. One group of boys was to be given the benefits of counseling while another group of boys matched on several variables was to remain untreated. Members of the two groups were selected from lists of names provided by various sources. School authorities nominated boys considered as difficult and troublesome as well as boys regarded as adjusted. Court records were examined for names of potential study subjects. Probation officers, police officers, social agencies, etc., were asked to submit names. Approximately 2,000 names were obtained. All boys who had passed their twelfth birthday during the period between referral and investigation were eliminated. Boys who could not be found or who were unavailable were also eliminated. The names of those remaining after this screening were submitted to three experts (not members of the project staff) for rating on an eleven-point delinquency probability scale. The rating process, which took fifteen months to complete, provided 782 candidates for the experiment. Two psy-

chologists were then asked to match one boy with another on such variables as health, intelligence, personality, home, neighborhood and delinquency prognosis. The toss of a coin determined which of the matched boys would be placed in the experimental group. Six hundred and fifty boys were selected and divided into two groups of 325 each.

Boys in the experimental group were assigned to the project staff. Staff work with such boys was begun in November, 1937, and was finished in May, 1939. Two years later it was found that case loads of 35 boys placed too great a burden on the counselors and in order to facilitate more effective work, boys believed to be in no need of service were dropped. This "retirement" (65 boys) plus the elimination of cases through death or mobility out of the project area (113 boys) and the manpower shortage produced by the war which necessitated discharging boys from care as they reached their seventeenth birthday (72 boys), resulted in subjecting the members of the group to varying periods of treatment. This in brief is the study organized to test Dr. Cabot's hypothesis.

The book contains two parts. Powers, the author of the first part, describes the project and the subjects chosen for treatment. Many phases of delinquency and its prevention in an urban setting are adequately discussed. In this part of the book the reader will also find a comprehensive treatment of the many problems which arose as the project evolved. The second part, written by Witmer, is concerned with the evaluation of the results of the experiment.

Before indicating the results of the experiment it seems desirable to describe briefly the personnel selected to implement Dr. Cabot's idea. The search for staff was begun in 1935 and in all over 250 persons were considered for the 10 counseling positions. In their search, the directors of the project were primarily interested in persons believed to possess intelligence, tact, unimpeachable character, and professional experience in dealing with people. Those selected were also to have faith in the objectives of the project. Formal education and training in professional social work were not considered a prerequisite if the candidate was "a warm, outgoing per-

son who had that vital spark so essential in human relationships" (page 92). Women, as well as men were considered since it was believed that the former would be particularly useful in dealing with younger boys. Of the ten persons chosen to begin the project four were women. A total of 19 different counselors were employed in the duration of the experiment; 15 men and 4 women. Of these 19, two had had experience as boys' workers, two were psychologists, one was a trained nurse, eight were professional social workers and six others had completed some of the academic requirements for a degree in social work.

The counseling staff of the project attempted to be of service to the boys in many ways. In general each counselor was expected to learn to know the boys assigned to him as completely as possible so as to aid the boy to make a more effective adjustment to changing life situations. To do this effectively the counselor needed to be intimately acquainted with each boy's assets and liabilities. But even more than this the counselors helped boys or members of their families to find employment, arranged for camp and summer placements, advised and counseled the boy's family in respect to his problems, procured professional services to remedy the boy's handicaps, taught and encouraged the boy to pursue hobbies and wholesome recreational activities, etc. Thus, as it may be seen, the counseling staff operated in many fields and stood ready to aid both the boy and his family to meet a variety of needs.

The results of the experiment indicate that little was accomplished. As a matter of fact, if we adhere strictly to the data presented, the differences in social adjustment of the boys in the experimental and control groups are insignificant. The services rendered by the project appear to be no more effective in achieving adjustment than the ordinary events in the lives of the boys. It seems apparent that delinquency cannot, on the average, be prevented by providing the services and counseling rendered by the project. All of this suggests that delinquency and maladjustment must be regarded as associated with a variety of combinations of psychosocial factors and any program intended to prevent such



deviations must provide different techniques to deal with such different combinations of factors.

This is an important book. Much too little has been done to put to a rigorous test the explicit or implicit assumptions that underlie much of what is done in social engineering. The experiment reported does this in a fashion that renders it an outstanding example of the best in social science research. Dr. Cabot's idea failed to produce hoped for results but the experiment designed to test the idea is a significant contribution to all of the social sciences.

Dr. Allport's foreword is an exceptionally well written preview of the study.

Elio D. Monachesi

Department of Sociology,  
University of Minnesota

*Personality: Symposia on topical issues*, Vol. 1, Nos. 3 and 4 (pp. 213-388). New York: Grune and Stratton, 1951.

Of these two numbers, the first contains 11 articles on Hypnosis and Personality and the second seven articles on Hypnotherapy. G. W. Williams' introductory article discusses some of the unsolved problems of hypnosis; here, as in many other places in this symposium, the controversial nature of nearly all topics in this field is emphasized. Guze writes on posthypnotic behavior and suggests that a standard experimental situation involving responses to posthypnotic suggestions might be useful as a diagnostic tool, since it would show how subjects handle impulses not congruent with their usual behavior. True and Stephenson present a very important research article on the EEG, pulse, and plantar reflex in age regression and induced emotional states, in which they confirm the recent finding that the Babinski reflex appears in subjects who are regressed to infancy, but they fail to find EEG changes. Harriman reports experiments on automatic writing, which resulted in very few conclusions. Loomis contributes a thorough survey of experiments from Bramwell to the present on space and time distortion in hypnosis. LeCron gives the results of an inquiry among

hypnotists, which shows that they are generally poor subjects. A short but vividly written article by Estabrooks discusses possible antisocial uses of hypnosis. Weitzenhoffer gives a survey of the major investigations of transcendence of normal voluntary capacities and concludes that such transcendence is fairly well established and that suggestions can cause alterations in nearly all organismic activities. As frequently happens when contributions are invited, a certain amount of recently published material is warmed up and served again. Christenson, for example, has published also in *Psychiatry* (1949) and in *Experimental Hypnosis* (edited by LeCron) articles on dynamics in hypnotic induction in addition to the one appearing here. A prospective rather than a retrospective look is, however, characteristic of the article by Kline on psychodiagnostic testing; of 15 articles cited eight are "in press."

In the number on Hypnotherapy there is a very useful introductory article by Schneck, which includes brief summaries of some of the literature. Watkins' "Hypnotherapy in the military setting" offers little to one who has read his book. Rosen's "Radical hypnotherapy of apparent medical and surgical emergencies" contains four full case reports and is largely new material. Kroger gives a very thorough account of personality dynamics and hypnosis in gynecology based upon Kroger and Freed's book. The articles by Raginsky (anesthesiology), Heron (dental uses), and Abramson (obstetrical uses) were to this reviewer tantalizingly brief and general in treatment, and his scanty information in these fields was not much increased by them. Fuller articles on these topics with more concrete descriptions of the procedures would have been welcome.

In spite of the question of multiple publication in a time when nearly all outlets for publication are crowded, this symposium is very valuable; it presents some new material, and its summaries and surveys and full bibliographies make it very useful for the student, investigator, and practitioner.

University of Kentucky

Frank A. Pattie

## New Books, Monographs, and Pamphlets

Books, monographs, and pamphlets for listing and possible review should be sent to Donald G. Paterson, Editor, Department of Psychology, University of Minnesota, Minneapolis 14, Minnesota.

- Maternal dependency and schizophrenia.* Joseph Abrahams and Edith Varon. New York: International Universities Press, 1953. Pp. 240. \$4.00.
- The design of social research.* Russell L. Ackoff. Chicago: The University of Chicago Press, 1953. Pp. 376. \$7.50.
- Personality fundamentals for administrators.* Chris Argyris. New Haven: Labor and Management Center, Yale University, 1953. Pp. 123.
- Roles and relationships in counseling.* Ralph H. Berdie, Editor. Minneapolis: University of Minnesota Press, 1953. Pp. 37. \$1.25.
- The social theories of Harry Stack Sullivan.* Dorothy R. Blitsten. New York: The William-Frederick Press, 1953. Pp. 186. \$3.50.
- Design for decision.* Irwin D. J. Bross. New York: The Macmillan Company, 1953. Pp. 276. \$4.25.
- Current theory and research in motivation.* Judson S. Brown, et al. Lincoln: University of Nebraska Press, 1953. Pp. 193. \$2.00.
- Professional problems in psychology.* Robert S. Daniel and C. M. Louttit. New York: Prentice-Hall, Inc., 1953. Pp. 416. \$5.50.
- Political community at the international level: problems of definition and measurement.* Karl W. Deutsch. Princeton: Princeton University Press, 1953. Pp. 71.
- Steps in psychotherapy.* John Dollard, Frank Auld, Jr. and Alice Marsden White. New York: The Macmillan Company, 1953. Pp. 222. \$3.50.
- Structure of human personality.* H. J. Eysenck. New York: John Wiley & Sons, Inc., 1953. Pp. 348. \$5.75.
- Farnum music notation test.* Stephen E. Farnum. New York: Psychological Corporation. Pp. 11.
- Symposium on fatigue.* W. F. Floyd and A. T. Welford, Editors. London: H. K. Lewis & Co. Ltd., 1953. Pp. 196. 24s net.
- Psychiatry and military manpower policy.* Eli Ginzberg, John L. Herma, and Sol W. Ginsburg. New York: King's Crown Press, 1953. Pp. 66. \$2.00.
- Sample survey methods and theory.* Volume I. Morris H. Hansen, William N. Hurwitz, and William G. Madow. New York: John Wiley & Sons, Inc., 1953. Pp. 638.
- Sample survey methods and theory.* Volume II. Morris H. Hansen, William N. Hurwitz, and William G. Madow. New York: John Wiley & Sons, Inc., 1953. Pp. 332. \$7.00.
- How to take a test.* Joseph C. Heston. Chicago: Science Research Associates, 1953. Pp. 47. \$4.00.
- Developmental psychology.* Elizabeth B. Hurlock. New York: McGraw-Hill Book Company, 1953. Pp. 556. \$6.00.
- Psychological reflections.* C. G. Jung. New York: Bollingen Series, 1953. Pp. 342. \$4.50.
- A court for children.* Alfred J. Kahn. New York: Columbia University Press, 1953. Pp. 359. \$4.50.
- Sexual behavior in the human female.* Alfred C. Kinsey, Wardell B. Pomeroy, Clyde E. Martin, and Paul H. Gebhard. Philadelphia: W. B. Saunders Company, 1953. Pp. 842. \$8.00.
- Hypnotism for professionals.* Konradi Leitner. New York: Stravon Publishers, 1953. Pp. 127. \$4.00.
- Films in psychiatry, psychology and mental health.* Adolf Nichtenhauser, Marie L. Coleman, and David S. Ruhe. New York: Health Education Council, 1953. Pp. 269. \$6.00.
- Applied imagination.* Alex F. Osborn. New York: Charles Scribner's Sons, 1953. Pp. 317. \$3.75.
- New light on dreams.* Max Serog. Boston: The House of Edinboro, Publishers, 1953. Pp. 159. \$3.00.
- Group relations at the crossroads.* Muzafer Sherif and M. O. Wilson, Editors. New



- York: Harper and Brothers, 1953. Pp. 379. \$3.50.
- Lawless youth.* E. A. Stephens. New York: Pageant Press, 1953. Pp. 315. \$3.50.
- The study of behavior.* William Stephenson. Chicago: University of Chicago Press, 1953. Pp. 376. \$7.50.
- Outline of executive development.* Lee Stockford. Pasadena: California Institute of Technology, 1953. Pp. 46. \$2.00.
- Living with a disability.* Eugene J. Taylor and Howard A. Rusk. New York: The Blakiston Company, Inc., 1953. Pp. 207. \$3.50.
- The work of a counselor.* Leona E. Tyler. New York: Appleton-Century-Crofts, Inc., 1953. Pp. 323. \$3.00.
- Recruiting the college graduate: A guide for company interviewers.* Richard S. Uhrbrock. New York: American Management Association, 1953. Pp. 31. \$1.25.
- How to help people.* Rudolph M. Wittenberg. New York: Association Press, 1953. Pp. 64. \$1.00.

## Personality Test Scores in the Management Hierarchy

Henry D. Meyer and Glenn L. Pressel

*Stevenson, Jordan & Harrison, Inc., Chicago, Illinois*

The primary purpose of this study was to obtain, if possible, an indirect but comprehensive industrial validation of the paper and pencil personality test developed by Stevenson, Jordan & Harrison psychologists for use as an interview aid in their work of appraising candidates and incumbents for management positions in industry. The basic concept implicit in the development of the test was that certain personality traits become increasingly desirable in incumbents and applicants as the positions bear increasing responsibility and relative status in the management hierarchy.<sup>1</sup> The most obvious test of validity of these traits would be to determine whether or not the people holding positions at different management hierarchy levels show differences in these same test traits and whether the differences show constant increments as the hierarchy levels increase.

Such a validation study does not make any discrimination between the competent and incompetent person at any given level. Rather, the assumption is made that some complex selective survival and elimination process is operating because consistently fewer persons achieve successively higher levels in the management hierarchy. If the personality traits tested are pertinent to such selectivity, that fact should be apparent in the distribution of trait scores at the various hierarchy levels.

This type of "selection" criterion was much more acceptable to the present authors than a criterion based on some authoritative group's judgments of the managerial competence of executives or managers. Not only did the former type of criterion eliminate the necessity of getting agreement of judges as to what

is managerial competence and how it is observed, but also it allowed the study to proceed around the design of a statistical analysis of previously obtained data from S. J. & H. files. The execution of the study therefore became a formal test of the hypothesis that there are trends in personality trait test scores as one proceeds from lower to higher level positions in the management hierarchy.

### Selection Procedures

The industrial management hierarchy was divided into five job levels with officers and general managers at the highest level and hourly rate workers at the lowest level. A total of 100 cases for each level except at the top<sup>2</sup> were selected from S. J. & H. personnel evaluation test files. Each case had been given, at the original time of testing, the improved Form B of the Employee Questionnaire, the S. J. & H. personality test.

*The Personality Test.* The Employee Questionnaire, known as the E. Q. Test, is a brief industrial personality test developed and described in the literature by previous S. J. & H. psychologists headed by H. F. Rothe (1, 3, 4) and subsequently improved by increasing the number of items from 50 to 75 and modifying the trait scoring keys according to the results of an item analysis. The tests were scored on seven trait keys with 8 to 12 items in each trait key and with simple scoring of items without weighting. These traits were objectivity, social dominance, drive, detail, emotionality, extraversion (sociability), and (poor) adjustment. Rothe (4) has discussed the definition of all of these trait terms except detail, which may be defined as the liking for detail in work, thought, and recreation; the desire to personally take care of all the details of projects in which one is involved. For each trait the mean score and the standard deviation for each of the five hierarchy levels was computed.

No claim is made that each trait is a pure, unitary factor. Rather, the definitions describe

<sup>1</sup> The word "management" is used broadly here to characterize all positions above the hourly rate level from foreman, engineer, salesman, or accountant, up to president.

<sup>2</sup> Only 57 cases were available from July 1949 to February 1952 which filled the requirements of being in the top category and taking the improved Employee Questionnaire, Form B personality test.



traits that are felt to be relevant to management success. The major intercorrelations among traits for 161 cases where objectivity is held constant at a median score are as follows: social dominance and extraversion,  $r = .78$ ; adjustment and emotionality,  $r = .67$ ; detail and emotionality,  $r = .66$ ; adjustment and detail,  $r = .52$ ; detail and drive,  $r = .46$ ; adjustment and social dominance,  $r = -.32$ ; emotionality and drive,  $r = .28$ ; adjustment and extraversion,  $r = -.26$ ; and extraversion and drive,  $r = .25$ .

A factor analysis<sup>3</sup> of the same data, i.e., with objectivity held constant, reveals two major clusters and one minor cluster. The first cluster is social dominance and extraversion; the second, detail, emotionality, and adjustment; and the minor one, drive.

*Categorizing the Hierarchy.* On the basis of the senior author's experience in consulting with industrial concerns at all levels of management, the hierarchy was broken down into five grades of job status according to job titles as follows:

I. President, Vice President, Treasurer, General Manager, General Sales Manager and Executive Engineer.

II. Works or Plant Manager, Sales Manager, Chief Engineer, Chief Industrial Engineer, Controller, Industrial Relations Director, Purchasing Director.

III. Production Superintendents, Industrial Salesmen, Sales Engineers, Department or Section Heads in Accounting, Industrial Engineering, Design Engineering, Inside Sales, Purchasing and Personnel.

IV. Production Foremen, Accountants, Design and Process Engineers, Time Study and Production Control Men, Sales Correspondents, Jr. Industrial Salesmen, Personnel Men.

V. Clerks and Factory Workers.

This breakdown of job status into hierarchy levels was reviewed and accepted by a supervising engineer of the S. J. & H. engineering staff as a rough approximation to the general trend in manufacturing industry insofar as one existed. This breakdown was used as a guide in sorting the actual cases into five hierarchy groups. While it was recognized that job titles are no guarantee of specific job content nor always a true reflector of the actual status of the job in the management, it was felt that such a breakdown came as close as was possible to a general criterion of management status. In any event, the selection of the cases followed procedures more elaborate than merely reading a job title as will be shown in the next section. These job titles were the convenient way of expressing distinctions in status that derived from the senior author's consulting experience and cannot be defended any further than that.

*Selection and Placement of the Cases.* In the selection of the cases, the major attempt was to obtain purity of hierarchical and occupational

classification with as wide a variation in occupation and company affiliation as permitted by the case history file. In studying cases for placement, the man's whole work history was reviewed. The criteria for selection and placement in a hierarchy were:

1. That the person be now, or last, employed at a job clearly recognized as belonging to a specific grade in the hierarchy or that his employment record indicate that he had consistently or steadily held such a job or jobs in the past.

2. That his employment record indicate a consistency of occupation and that he had proceeded through a normal job succession up to the job according to which he was classified.

3. That if the occasion for the testing was to apply for a position, the job applied for be at the hierarchy level indicated by his previous employment.

In reviewing the cases, the senior author frequently was able to bring a personal knowledge of company size and organization structure to bear upon the information provided in the job history. Also, since all of the cases had to have been tested since July of 1949, to have taken the improved E. Q. Form B personality test, the senior author had interviewed the majority of the cases himself regarding their job duties and histories. As a result of this knowledge of company and job, a more consistent selection of typical cases for each grade was obtained than could have been obtained from job titles alone.

No attention was paid to test results in selecting cases, nor was any consideration given to whether the person was appraised as superior, average, or inferior. In fact, many of the cases had to have their personality tests rescored on the improved key<sup>4</sup> after they had been chosen for the study.

The attempt to secure 100 cases for each of the five categories resulted in some stretching of the criteria in categories I and II of officers and second level executives where for the last few cases some men were chosen where the previous employers were not well known and where the job applied for was lower than the category in which the man was placed by reason of previous employment.<sup>5</sup>

Also it should be noted that there was no attempt to control company size or to modify the status of a job according to the size of the company. Companies of all sizes are listed in the employment histories of the cases selected. How-

<sup>4</sup> The scoring key was revised following the item analysis.

<sup>5</sup> A detailed summary of the present or most recent employment of all cases chosen for each hierarchy category has been deposited with the American Documentation Institute. Order Document No. 4191 from the ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress, Washington 25, D. C., remitting in advance \$1.25 for 35 mm. microfilm or \$1.25 for 6 x 8 in. photocopies. Make checks payable to Chief, Photoduplication Service, Library of Congress.

<sup>3</sup> The authors are indebted to Vernon Keenan for making this factor analysis.

Table 1

E. Q. Test Trait Score Means and Sigmas  
According to Hierarchy Level

E. Q. Trait	Hierarchy Level*									
	Total Group		I		II		III		IV	
	M	S.D.	M	S.D.	M	S.D.	M	S.D.	M	S.D.
Objectivity	4.4	1.9	4.3	1.9	4.2	2.0	4.2	1.9	4.6	1.8
Social Dominance**	6.7	2.2	7.2	2.0	7.5	2.1	6.7	2.2	6.4	2.3
Extraversion	3.9	1.8	3.7	2.0	3.9	1.7	4.0	1.9	4.0	1.7
Drive	5.9	1.8	5.6	1.7	5.8	1.9	6.0	1.9	6.2	1.7
Detail**	4.3	1.9	3.2	1.6	3.9	1.9	4.4	1.9	4.7	2.0
Emotionality**	4.0	2.3	3.0	1.9	3.7	2.3	3.7	2.3	4.3	2.5
Adjustment (poor)**	3.6	2.1	2.8	1.8	3.2	1.8	3.7	2.1	3.8	2.2

\* N = 100 for levels II, III, IV, and V. N = 57 for level I.

\*\* Significant at 5% level for single classification F test.

ever, most of the companies in which the persons were presently employed or seeking employment were medium-sized companies, or medium-sized plants of large companies. Typical employment would be between 500 and 1,000 for the smaller plants in the group and between 2,000 to 3,000 for the larger plants.

*Statistical Procedures.* The statistical procedures of the present study were based on the results of a pilot study of 200 cases chosen at random. The pilot study was a preliminary test of the hypothesis of a relationship between personality trait scores and management hierarchy levels in order to determine whether or not the hypothesis had sufficient merit to warrant a full scale major study.

The results of the pilot study indicated that trends for the E. Q. test traits by hierarchy level probably existed for the social dominance, detail, and emotionality traits and that there might well be continuous increment trends. An analysis of the differences between the means within the traits showing trends indicated that it would be necessary to have 100 cases at each of the compared hierarchy levels in order that mean differences of the magnitude observed be statistically significant. Hence, the major study was done with 100 cases in each hierarchy level except level I, for which only 57 cases were currently available. All subsequent results are from the major study.

### Results

#### *The Test for Trait Score Trend Validity.*

A single classification F test for hierarchy alone was used to determine the validity of the hierarchy trend for each trait and a *t* test was used to establish the validity of differences between trait score means at the extremes of the hierarchy for each trait.

These tests demonstrated significant hierarchy trends for social dominance, detail, emotionality, and adjustment, but failed to reveal significant hierarchy trends for objectivity, extraversion, and drive. Table 1 shows the trends in trait score means by hierarchy level for the major study.

*The Test for the Independence of Hierarchy Trends.* The establishment of hierarchy trends in several trait scores did not prove that hierarchy produced them independently of other variables. There could be other variables producing trait score trends which are associated with differences in hierarchy level. Age and education are certainly associated with differences in hierarchy level and might also produce trait score trends. Selective sampling by occupation might have occurred such that the hierarchical trends observed might have been due to trait score differences related to occupations rather than to hierarchy. Also, the present study and previous studies (4) indicated that differences in objectivity trait scores were known to produce differences in the four trait scores having hierarchy trends to as great an extent as hierarchy levels, particularly in the emotionality and adjustment traits. Fortunately, the objectivity score means were approximately the same for all five hierarchy levels in the present study.

The first step in testing the independence of hierarchy as the trend producing variable was the analysis of the data to see if there



Table 2

E. Q. Test Trait Score Means and Sigmas  
According to Age Level

E. Q. Trait	Age Level							
	Years 20-30 N = 87		Years 30-40 N = 171		Years 40-50 N = 152		Years 50+ N = 39	
	M	S.D.	M	S.D.	M	S.D.	M	S.D.
Objectivity	4.7	2.0	3.5	2.1	4.1	1.9	4.3	1.8
Social Dominance	6.8	2.4	6.6	2.4	7.0	2.2	6.6	1.8
Extraversion	4.2	1.7	4.2	2.8	3.7	1.7	3.4	1.6
Drive	6.2	1.6	5.9	1.9	6.0	2.0	5.5	1.6
Detail	4.8	1.8	4.4	2.0	4.1	1.9	3.9	1.9
Emotionality	4.7	2.4	4.1	2.5	3.5	2.2	3.7	2.3
Adjustment (poor)	3.8	2.2	3.8	2.6	3.4	2.0	3.7	2.2

were observable trends in the alternative variables noted, i.e., age, education, and occupation. These results are shown on Table 2 for age, Table 3 for education, and Table 4 for occupation.

This analysis indicated that the hierarchy trend for adjustment trait scores might have been due to education alone and that the trends for detail and emotionality trait scores might have been due to both education and age. Occupational differences were too small to admit of much possibility for causing the observed hierarchy trends.

The second step in testing the independence of hierarchy as the trend producing variable was a double classification F test analysis of variance pairing hierarchy with each of the variables—age, education, occupation, and objectivity.<sup>6</sup> This would show, if the data were adequate, whether trait score trends for hierarchy were independent of trait score trends for these four variables, co-existent with them but still independent, or interacting with them to produce the over-all effect labeled "hierarchy trend."

Several difficulties occurred in the execution of this procedure because of the unequal distribution of two of the four variables throughout the hierarchy. Only two graduations of age could be used—30-40 years and 40 years and up, because no cases under 30 years were in hierarchy levels I and II, and

few in III. Also, only three hierarchy levels could be paired with occupation because occupational specialization is infrequent in hierarchy I, which is composed of cases in general management at the officer level and because at level V, the lowest in the hierarchy, the five occupational groups of production supervision, sales, accounting, design engineering, and industrial engineering, give way to hourly rate jobs with skill, craft, clerical, or service classifications.

This shrinkage in the number subclassification for one of the paired variables combined with the shrinkage in the number of cases in each cell because of the failure to use all of the data in the age and occupation pairings with hierarchy made it much more difficult to establish statistically significant independent trends with these two variables than with the other two variables, education and objectivity. As a result of these technical difficulties, the double classification analysis of variance technique used in this study is thought to be valid only on the positive side. That is, where the double classification F test shows the independence of the hierarchy trend from one of the four paired variables for a given trait, that can be accepted as proof positive of independence. But where independence of hierarchy trend is not shown by this technique, the matter is still open to subsequent proof or disproof using more cases and more numerous subclassifications of the paired variables.

The double classification F test pairing

<sup>6</sup> Because the subclasses contained widely differing numbers, the "disproportionate subclass method" for treating data classified in unequal numbers of items was used (4, 235-240).

Table 3

E. Q. Test Trait Score Means and Sigmas  
According to Education Level

E. Q. Trait	Education Level*					
	High School N = 179		2 Years College N = 85		College Graduate N = 193	
	M	S.D.	M	S.D.	M	S.D.
Objectivity	4.4	1.8	4.7	1.9	4.3	1.9
Social Dominance	6.5	2.3	6.4	2.1	7.2	2.1
Extraversion	3.8	1.7	3.7	1.7	4.2	1.8
Drive	6.2	1.9	5.9	1.5	5.7	1.7
Detail	4.8	1.8	4.4	1.8	3.9	1.8
Emotionality	4.6	2.3	3.7	2.5	3.6	2.2
Adjustment (poor)	4.4	2.2	3.4	2.0	3.0	2.0

\* High School = High School graduation or less; 2 Years College = 1 or 2 years college; and College Graduate = 3, 4 or more years of college.

hierarchy in turn with age, education, objectivity, and occupation failed to demonstrate any significant trend independence for drive and extraversion. Adjustment was shown to have significant trend independence at the 5% level or better when hierarchy was paired with age, education, and objectivity; emotionality and social dominance had significant trend independence when paired with age and education; detail had significant trend independence when paired with education and objectivity.

The complete absence of independent trends for the occupation variable when paired with

hierarchy is probably due to the small variance of trait scores among occupations. The complete absence of independent hierarchy trends when paired with occupation must be discounted as possibly due to a limited number of hierarchy subclasses in this pairing, classes II-IV. Also, the absence of some expected independent trends for the age, education and objectivity variables must be discounted as possibly due to the limited number of subclassifications for these variables—2 for age and 3 for education and objectivity as compared with 5 for hierarchy. Keeping the above technical limitations in mind, cer-

Table 4

E. Q. Test Trait Score Means and Sigmas  
According to Occupation

E. Q. Trait	Occupation*									
	A N = 63		B N = 36		C N = 35		D N = 32		E N = 84	
	M	S.D.	M	S.D.	M	S.D.	M	S.D.	M	S.D.
Objectivity	4.1	1.8	4.8	1.9	4.6	1.9	3.8	2.1	4.3	2.0
Social Dominance	6.3	1.9	6.7	2.3	7.1	2.1	6.7	2.2	6.7	2.4
Extraversion	3.9	1.7	3.2	1.7	4.1	1.8	3.9	1.6	4.8	1.8
Drive	6.0	2.2	5.7	1.8	6.3	1.7	5.3	1.8	6.3	1.7
Detail	4.4	1.7	4.5	2.2	4.3	1.7	4.3	1.8	4.3	2.1
Emotionality	4.0	2.3	4.1	2.4	4.1	2.2	3.6	2.3	3.9	2.5
Adjustment (poor)	4.1	2.0	3.8	2.3	3.7	2.1	3.3	1.7	3.9	2.2

\* A = Production Supervisors; B = Design Engineers; C = Salesmen; D = Accountants; E = Industrial (Production) Engineers.



tain statements may be made about the observed valid trends in E. Q. test trait scores previously presented in Table 1.

### *Summary of Primary Results*

1. The trend for higher social dominance trait scores as the hierarchy ascends is: (a) independent of age; (b) independent of education; (c) not proven to be independent of objectivity; and (d) not proven to be independent of occupation.

2. The trend for lower detail scores as the hierarchy ascends is: (a) not proven to be independent of age; (b) the result of interaction of hierarchy and education even though there is some other quantitative degree of the hierarchy trend which is independent of education; (c) independent of objectivity; and (d) not proven to be independent of occupation.

3. The trend for lower emotionality scores as the hierarchy ascends is: (a) independent of age; (b) independent of education; (c) not proven to be independent of objectivity; and (d) not proven to be independent of occupation.

4. The trend for lower, i.e., better adjustment scores as the hierarchy ascends is: (a) independent of age; (b) probably independent of education (very close to 5% level of confidence) although there is a similar reduction trend with increasing education that is independent of hierarchy; (c) a result of the interaction of the hierarchy and objectivity variables even though there is some other quantitative degree of the hierarchy trend which is independent of objectivity and some objectivity trend probably independent of hierarchy (very close to 5% level of confidence); and (d) not proven to be independent of occupation.

*Secondary Results.* A number of trends in trait scores were observed for the variables of age, education, objectivity, and occupation alone. Of these, only age was tested for validity of trend by a single classification F test analysis of variance. This was done because the double classification F test pairing age with hierarchy was felt to be inadequate because the full age range of the data could not be used. The single classification F test for age alone showed a valid trend, at better than

the 5% level of confidence, for detail trait score means of successively older age groups to decline. The previously observed trends for extraversion (sociability), emotionality and drive trait scores to decline with increasing age (Table 2) were not marked enough to prove themselves valid at the 5% level of confidence.

Increasing amounts of formal education gave trait score trends of lower (poor) adjustment, emotionality, detail, and drive trait score means (Table 3). These trends were not tested for validity by the single classification F test because the double classification F test pairing education and hierarchy was thought to be adequate. As indicated previously, only the trend for adjustment scores to decline with education proved valid at the 5% level of confidence.

Occupational differences in trait score means occurred but could not be called trends (Table 4). The differences were not great enough and the number of cases in each occupational group was too small to establish their validity by statistical procedures. Salesmen and industrial engineers were highest in traits of extraversion (sociability) and drive. Design engineers were lowest in extraversion and highest in objectivity. Accountants were lowest in objectivity and (poor) adjustment, the latter probably because of the former. Production supervisors were highest in adjustment. There were no marked differences among the five occupational groups segregated in the trait scores of detail or emotionality.

Decreasing objectivity trait score groups gave trait score trends of higher extraversion and lower detail, emotionality, and (poor) adjustment trait scores. Only the (poor) adjustment trait score trend for objectivity differences proved valid in the double classification F test pairing objectivity and hierarchy. The use of a larger number of objectivity subgroupings and more cases would be required to clarify whether the traits of dominance, detail and emotionality also have valid trends with objectivity.

### *Discussion*

Inasmuch as two of the failures of traits to establish the independence of their hierarchy

trends occurred when hierarchy was paired with objectivity, it is of utmost importance to recognize that objectivity trait score means were practically constant for all five grades of the hierarchy. Hence, while it can be concluded that hierarchy trends will not occur in emotionality or dominance trait scores when dealing with only high or low objectivity score groups, it can be said that these effects will cancel each other out in a randomly selected sample that gives a normal distribution of objectivity scores. Therefore, in the present study, where a normal distribution of objectivity scores occurred at all hierarchy levels, emotionality and dominance trait score trends appeared which were not due to objectivity differences among the hierarchy grades.

The hierarchy trend for detail trait scores cannot be separated from the age variable at the present time because of the lack of younger people in the upper hierarchy levels. However, age is also closely related to administrative and managerial experience in the present sample and such experience could be the true variable associated with the detail trait hierarchy trend. A control study, keeping age constant at 35 to 45 years with the  $N$  at each hierarchy level ranging from 27 to 67 cases gave a consistent hierarchy trend for detail of about the same magnitude as in the major study (Table 1) in which age was uncontrolled. Hence the total evidence favors the independence of the hierarchy trend for the detail trait from the age variable, if not from managerial experience.

It is also apparent that occupational influences on trait scores overlie hierarchy influences and in a few cases may exceed them. For example, junior salesmen had higher dominance scores than sales managers in our small sample of the sales occupation. Furthermore, the fact that hierarchy differences in trait scores were greater than occupational differences points up the fact that vocational guidance for adults, relative to tested personality traits, has a hierarchy level dimension which may be more discernible than the occupational dimension.

Extremely interesting to the authors is the fact that two of the four traits showing valid hierarchy trends, i.e., detail and emotionality,

showed the least differences among occupational groups of all seven traits. Only in positions deemed administrative, such as general managers, works managers, and industrial relations directors did a sharp reduction in emotionality and detail trait score means become apparent. This suggests the hypothesis that hierarchical differences are primarily differences in the breadth and generality of administrative responsibilities. The hierarchy, according to this hypothesis, proceeds from specific occupational activities to the administration of specific occupational activities, to the administration of more diverse and more generalized occupational activities; and rising in the hierarchy is favored by personality traits suitable in degree and kind to such administrative responsibilities.

As a precaution against overgeneralizing the results of this study, it should be remembered that at every hierarchy level there was a nearly normal distribution of scores for every trait and that the observed trait trends were only small changes in the central tendencies of these distributions. That is why it was so difficult to establish the validity of these trends. It took five hierarchy levels with approximately 100 cases in each to do it. The wide dispersion of scores could be due to the fact that there are many other probable determiners of hierarchy-climb survival or achievement than personality trait scores. Intelligence, experience, education, political skill, competition, motivation, values, etc., are other variables that come to mind. Considering these many probably contributing variables, it is remarkable that a brief pencil and paper personality test could show consistent and valid hierarchy trends in four of its seven trait scores. It should also be noted that of the two traits with the highest intercorrelation, social dominance and extraversion (sociability)  $r = .78$ , only one, social dominance, showed a hierarchy trend. This throws doubt on the validity of Ellis' (2) implied criticism that personality inventories with overlapping traits are undesirable.

Also, it should be noted with caution that the present study does not offer any direct evidence as to whether possession of these "trend" trait scores on the "high" side of the distribution at a single level of the hierarchy



indicates that their possessors are more competent in their jobs than people on the "low" side within that same hierarchy level. Rather, the evidence is all indirect and follows a "survival" concept based on the belief that there is a progressively more stringent selection for fewer and fewer jobs as the hierarchy ascends. Since the variance in four trait scores has been demonstrated to have hierarchy "survival" value by this study, it can be concluded that this study has been, to some degree, successful in indirectly validating S. J. & H.'s E. Q. Personality test for use as one tool among several in their appraisal of candidates and incumbents for management positions. Also, in a limited way, it has contributed in the broader task of isolating the characteristics of industrial managers. It remains for a future study to determine whether there are additional hierarchy trend traits and whether all trend traits "develop" in their possessors with job experience or exist full blown from early adulthood.

#### Summary

The traits of (poor) adjustment, emotionality, detail and social dominance as measured by Form B of the Employee Questionnaire, a brief industrial personality test developed by Stevenson, Jordan & Harrison psychologists (1), were found to have valid management hierarchy trends. The traits of extraversion (sociability), drive, and objectivity did not have valid hierarchy trends. There was no rating criterion of success or failure for the cases studied. Rather, current achieved position in the hierarchy was the implied criterion since the cases studied held jobs at a particular level in the hierarchy.

The criterion of validity for trend was a single classification analysis of variance of trait scores for the five hierarchy categories giving an F ratio at the 5% level of confidence or better. Also, validity was demonstrated by a *t* test of the significance of differences between trait score means of the top and bottom hierarchy categories giving a *t* ratio at the 5% level of confidence or better. The valid trends were such that detail, emotionality and (poor) adjustment trait score means decreased at each successively higher level of the hierarchy while social dominance

trait score means increased as the hierarchy levels became successively higher.

The industrial management hierarchy was divided into five levels with company officers and general managers at the top level and hourly rate employees at the bottom level. One hundred cases at each of the five levels except the top level, which had 57, were utilized in the study.

The valid hierarchy trends for the four traits mentioned were found by a double classification analysis of variance technique to exist independently of the variables of age, education, and objectivity at the 5% level of confidence or better with the following exceptions. The detail trait trend was not proven to be independent of age. The social dominance and emotionality trait trends were not found to be independent of objectivity. The latter exception was not held to be a major defect in demonstrating hierarchy trends because objectivity trait score means were practically equal at all five hierarchy levels.

Miscellaneous trait trends were observed for the variables of age, education, objectivity, and occupation alone. But these were not a major aspect of study and were not tested statistically. Hierarchy differences in personality trait scores were generally greater than occupational differences.

For all traits there was a substantial and normally distributed dispersion of scores around the mean at every level of the hierarchy indicating the probable participation of many variables in addition to personality test scores in determining the hierarchy levels of the cases studied.

Received May 27, 1953.

#### References

1. Carr, E. R., and Rothe, H. F. Validity of an objectivity key on a short industrial personality questionnaire. *J. appl. Psychol.*, 1950, 34, 178-181.
2. Ellis, A. Recent research with personality inventories. *J. consult. Psychol.*, 1953, 17, 45-49.
3. Mitchell, M. B., and Rothe, H. F. Validity of an emotional key on a short industrial personality questionnaire. *J. appl. Psychol.*, 1950, 34, 329-332.
4. Rothe, H. F. Use of an objectivity key on a short industrial personality questionnaire. *J. appl. Psychol.*, 1950, 34, 98-101.
5. Snedecor, G. W. *Statistical methods* (4th Ed.). Ames, Iowa: Iowa State College Press, 1946.

## Temperament Measures in Industrial Selection<sup>1</sup>

Frederick Herzberg

*Psychological Service of Pittsburgh and University of Pittsburgh*

Psychologists have reached an advanced stage in the development of test measures which can be applied to the problems of industrial selection. The least dependable of these measures lies today in the area of personality and temperament assessment. One of the major reasons for the wariness with which industrial psychologists approach temperament inventories is the transparency of such tests and their corresponding amenability to faking or pointing answers to achieve a desired result. Many studies have demonstrated the possibility that this faking can occur (1, 2, 3, 4, 5). These studies, however, have generally been based on artificial situations in which college students have been instructed to attempt such faking.

One may ask, as Guilford does, in his manual for the Guilford-Zimmerman Temperament Survey, whether such faking or pointing occurs in the actual situation. And if faking does occur, to what extent does it negate or limit the use of the test for industrial selection purposes?

Two hypotheses were examined in this study relating to the existence in the employment situation of such manipulation of test responses. The first hypothesis is that the distribution of the Guilford-Zimmerman Temperament scores will be significantly higher for persons tested in the industrial situation than the distribution of scores for either college students or of clients seen for vocational guidance.

It is suggested here that these three groups have three different motivations for taking this test. The motivation for the industrial group is to get a job or to get promoted. The motivation of the vocational counseling subjects ostensibly is to gain information about their abilities and job opportunities, while the college students' reason for taking the test is

basically an academic one of pleasing the instructor or participating in an experiment.

The second hypothesis is that in industrial testing, where faking is expected to occur, the educational level of the examinees will affect the extent of such faking; i.e., the higher the education the higher will be the score distributions. This is suggested by the fact that the higher educational groups have more general intelligence to understand the implications of the items and more test sophistication from their longer academic experience.

The Guilford-Zimmerman test was chosen for this study because it is one of the most widely used personality inventories of the non-psychiatric type. Its avoidance of psychiatric terminology and goals make it more applicable in industrial personnel work.

### Method

*Population.* The industrial group (those tested for employment, promotion, or company personnel survey purposes) consists of a total of 924 cases, of which 338 are college graduates, 128 have had 1-3 years of college education, 353 are high school graduates, and 105 have only elementary school education.

The self-referral group (vocational guidance clients) contains 94 college graduates and 56 high school graduates.

The college group (University of Pittsburgh students in Introductory Psychology classes) consists of a total of 109 students approximately equally distributed among the four years of freshman to senior.

All subjects are males.

*Analysis.* Frequency distributions and basic distribution statistics were computed for each of the "motivation" and educational groups. There was considerable skewness in many of the distributions with corresponding unequal variability between the comparison groups. All distributions were unimodal and plots showed that the differences between groups lay in higher scores for one distribution as

<sup>1</sup> This research was supported by a grant from the Buhl Foundation.



Table 1

Summary of Significant Differences Between Means of Guilford-Zimmerman Scales for Groups Studied

Scales	Industrial College Graduates vs. Industrial High School Graduates	Self-Referral College Graduates vs. Self-Referral High School Graduates	Pitt Freshmen and Sophomores vs. Pitt Juniors and Seniors	Total Pitt Sample vs. Industrial Non-College Graduates	Industrial College Graduates vs. Self-Referral College Graduates
Gen. Activity					**
Restraint				**	
Ascendence	**			*	**
Sociability	**			**	**
Emot. Stability	**			**	**
Objectivity	**	**		**	**
Friendliness				**	**
Thoughtfulness				**	
Personal Rel.	**	**		**	**
Masculinity	**		**		

\* Difference between means significant at .05 level of confidence.

\*\* Difference between means significant at .01 level of confidence.

opposed to the other. The differences between the means of the various groups were tested for significance by Student's "t" ratio. The .01 level of confidence was accepted.

There were significant age differences between some of the groups but this was proved not to be a pertinent variable in this study.

### Results

Table 1 presents a summary of the significant differences found between the means of the groups studied on the ten Guilford-Zimmerman scales.

A comparison on the basis of education within the industrial group shows higher mean scores for each scale on the Guilford-Zimmerman with increasing education from grammar school through high school to college graduation. Scales for which the differences between high school and college education are significant at the .01 level of confidence are *Ascendence*, *Sociability*, *Emotional Stability*, *Objectivity*, *Personal Relations*, and *Masculinity*. All these differences with the exception of *Masculinity* are significant at beyond the .001 probability level.

Two differences at the .01 level (*Objectivity* and *Personal Relations*) occur between college graduate self-referrals and the high

school graduate self-referral sample. Only for the *Masculinity* scale is there a significant difference between University of Pittsburgh freshmen and sophomores and Pittsburgh juniors and seniors. All three differences are again in the direction of larger means for the higher education groups.

In order to compare the industrial population with an academic motivation group the norms provided by Guilford could have been utilized. However, since our industrial subjects are from Western Pennsylvania and the manual norms are based upon California college students, it was decided to gather norms on an equivalent Pittsburgh college population. University of Pittsburgh norms are found to be essentially similar to those reported by Guilford with the exception of a higher *Masculinity* score for Pitt students. The Pittsburgh college group was then compared with non-degree college education industrial cases. This latter group was chosen for comparison with the Pitt students in order to equate for the education level which was found to be of significance for the industrial subjects. The norm values for the non-degree college industrial sample are found to be approximately midway between the norms for high school graduates and college graduates.

Significant differences at the .01 or better level of confidence differentiate these two groups in favor of the industrial population on the *Restraint*, *Sociability*, *Emotional Stability*, *Objectivity*, *Friendliness*, and *Personal Relations* scales. A higher *Ascendancy* mean was significant at the .05 point.

Comparing the college graduate industrial group with college graduate counseling clients, we find the means of *General Activity*, *Ascendancy*, *Sociability*, *Emotional Stability*, *Objectivity*, and *Personal Relations* scales all to be significantly different. These differences are once more in the predicted direction of higher scores for the industrial cases.

#### Discussion

These results support both of the hypotheses stated in the introduction. The industrial population for equivalent educational level have higher means on most of the scales of the G-Z than do corresponding academic and counseling client samples. In addition, the educational differences occurred primarily with the industrial samples. The hypothesis that faking or pointing of personality tests does actually occur in the industrial situation is well substantiated by these data; first, by their higher scores, and second, by the reinforcing of this finding with the educational differences obtained. It seems reasonable therefore to conclude that clients for employment or promotion do fake their test re-

sponses and this occurs to a greater extent at the higher educational levels.

As to the question raised in the introduction regarding the significance of such pointing on the usefulness of the test, one need only examine norms based upon a college graduate industrial sample. For the *Sociability* scale, the median will fall at a score of 25 on a 30 item scale, i.e., one-half of the group will achieve scores of five-sixths or more of the possible number of items included in that area. Similar results are found for the *Emotional Stability*, *Objectivity*, and *Personal Relations* scales.

Perhaps the nature of the distribution of G-Z scores which are obtained in employment testing is best illustrated by presenting the percentile ranks of scores for the groups studied which are equivalent to the medians on the manual norms. These percentile ranks appear in Table 2.

The median scores, for example, on the published norms for the *Emotional Stability* and *Personal Relations* scales fall at the fifteenth percentile when computed from a sample of industrial college graduates. The other scales show similar discrepancies in the median values. The equivalence of the Pitt sample to Guilford's California college population is shown in the last column of Table 2.

With such extreme "piling-up" it is difficult to conceive of the meaning of a high score on these scales, much less to utilize

Table 2

Percentile Ranks\* of Scores for the Groups Studied Which Are Equivalent to the Medians on the Manual Norms

Scales	Industrial College Graduates	Industrial High School Graduates	Self-Referral College Graduates	Self-Referral High School Graduates	Pitt Students
G	40	40	65	55	50
R	30	30	40	50	45
A	25	40	50	60	40
S	20	30	40	45	40
E	15	25	40	60	50
O	20	45	40	60	55
F	35	35	45	50	55
T	40	40	40	50	45
P	15	30	30	55	55
M	40	50	55	50	55

\* To the nearest fifth percentile.



them. When one considers the curvilinear use of these tests, as recommended by Guilford, he would have to reject half the applicants on those scales where a high score is considered a drawback.

These extreme results apply mostly to the use of the test with a college graduate industrial population. But this is just the population wherein the need for such a temperament evaluation is greatest. The top level jobs involving supervision and personal relations usually are held by college graduates, increasing the need of some assessment of their personality characteristics.

Received May 14, 1953.

## References

1. Cofer, C. N., Chance, June, and Judson, A. J. A study of malingering on the MMPI. *J. Psychol.*, 1949, 27, 491-499.
2. Hunt, H. F. The effect of deliberate deception on Minnesota Multiphasic Personality Inventory performance. *J. consult. Psychol.*, 1948, 12, 396-402.
3. Kimber, J. A. M. The insight of college students into the items on a personality test. *Educ. psychol. Measmt.*, 1947, 7, 411-420.
4. Longstaff, H. P., and Jurgenson, C. E. Fakability of the Jurgenson Classification Inventory. *J. appl. Psychol.*, 1953, 37, 86-89.
5. Wesman, A. G. Faking personality test scores in a simulated employment situation. *J. appl. Psychol.*, 1952, 36, 112-113.

## A Validation Study of the Worthington Personal History Blank

John G. Clark and W. A. Owens

*Iowa State College, Ames, Iowa*

The Worthington Personal History Blank (hereafter,—PH) consists of an unstructured 4-page application blank which is used as a projective technique in industrial selection. Some evidence for the validity of PH has appeared in the form of "testimonials" from satisfied users. Most of these have been favorable. Somewhat more empirical evidence has come from Worthington in his doctoral dissertation (5) and in a recent article (3). The former offers evidence of a favorable comparison (approximately 87% agreement) between PH analyses and psychiatric diagnoses for ten V.A. Mental Hygiene Clinic patients. The latter article indicates that PH was useful in predicting effectiveness of salesmen for a light manufacturing company, as indicated by biserial correlation coefficients of .34 with tenure and .31 with sales volume. Swint and Newton (4) reported that in the prediction of supervisory potentiality, the PH was accurate in 85% of the cases.

Since a single PH analysis costs in the neighborhood of \$40.00, its use would be justified only if its efficiency were considerably greater than that of conventional, less expensive instruments. It is, therefore, the purpose of this study to compare PH and objective tests with respect to their relative efficiency in predicting associates' ratings in an industrial situation.

### Method

The subjects of the present study were 47 employees of an Iowa publishing company. They were originally selected by the employment manager as having rather distinctive and unusual personalities lending themselves to easy PH diagnosis and to simple inspectional checks on the accuracy of the protocols. In order to make possible a more exacting test of the potential value of the instrument, the problem was subsequently presented to the Department of Psychology at Iowa State College as a possible thesis

project. The department approved it as such, and it was decided that the most convenient procedure for evaluation would be to compare the validity of the PH, against a criterion of associates' ratings, with the validity of certain standardized tests, against the same criterion.

In addition to PH analyses for each subject, percentile ranks were available on speed and power measures of intelligence (The Wonderlic Personnel Test and The Personnel Laboratory's Employment Test), on the Thurstone Temperament Schedule, and on the traits "Dominance" and "Self-Sufficiency" from the Bernreuter Personality Inventory.

A five step criterion rating scale was constructed, the traits included being selected on the bases of ease of rating and commonality with both PH and test results, particularly the former.

PH reports were transformed into quantitative terms by five experienced psychologists. These judges decided, on the basis of PH reports, whether a given subject should be classified as "high" (+) or "low" (−) with respect to each of the traits under consideration, and the score assigned reflected the degree of their agreement. Literally, a six point scale was provided, ranging from five +'s to no +'s. It was unnecessary to perform this operation on the PH estimates of intelligence, since these were reported in terms of estimated Wechsler-Bellevue intelligence quotients.

The criterion ratings were made by two raters per subject. Since the subjects were scattered throughout the company concerned, it was impossible to obtain more ratings per individual or to have each rater rate all of the subjects; and the ratings were, therefore, simply pooled.

Since the test scores were reported *only* in centile ranks and could not be assumed to be normally distributed, the correlation among the variables was estimated in terms of con-



Table 1

Agreement among Judges on Quantification  
of PH Reports

Traits	Degree of Agreement <sup>a</sup>	Degree of Agreement <sup>b</sup>
Activity	68%	87%
Vigorous	45%	74%
Impulsive	45%	74%
Dominant	42%	91%
Stable	38%	59%
Sociable	42%	84%
Reflective	19%	57%
Self-sufficiency	38%	78%
Job-effectiveness	70%	89%
Promotion possibilities	42%	72%
Adjustment to others	36%	74%

<sup>a</sup> Per cent of cases in which all five judges agreed.

<sup>b</sup> Per cent of cases in which four of the five judges agreed.

tingency coefficients.<sup>1</sup> The following comparisons were made: Comparison I: PH vs. ratings; Comparison II: Test results vs. ratings; Comparison III: PH vs. test results; and Comparison IV: PH vs. ratings as compared with test results vs. ratings.

The significance of the difference between the two series of contingency coefficients in Comparison IV was tested through the use of a randomization test (1).

### Results

The degree of agreement among the judges who quantified the PH reports is shown in Table 1.

Table 2 shows the retest reliabilities of the ratings.

Tables 3, 4, 5, and 6 show the results of the four comparisons. The only significant relationship found in the first three comparisons was that between PH and the Thurstone Temperament Schedule on the trait "Sociable." This coefficient was significant at the .01 level of confidence.

Comparison IV, the crucial one (Table 6) shows consistently higher coefficients of relationship between objective test results and ratings than between PH and ratings. The randomization test indicates a probability of

<sup>1</sup> Test scores were grouped by deciles.

Table 2

Retest Reliabilities of Ratings

Note: 5% level,  $r = .29$ ; 1% level,  $r = .37$ .

Trait	$r_{11}$
Activity	.68
Decisiveness	.46
Dominance	.63
Personal Adjustment	.90
Sociability	.62
Job-effectiveness	.79
Promotion possibilities	.93
Adjustment to others	.91

.06 that five such differences in the same direction could occur by chance.

### Discussion

The apparent, consistent superiority of the objective tests, as indicated in this investigation, constitutes damaging evidence as to the usefulness of PH. These results, while not in accord with those of the previously mentioned studies of the PH, do tend to follow the pattern of dubious or negative results found in validation studies of other projective techniques (2).

Although not of major importance to the present study, the lack of significant relationship to the criterion on the part of both structured and unstructured techniques is of interest. There are certain methodological problems involved which may, in part, account for this lack of relationship. These are centered around: (a) the criterion ratings; (b) the quantification of PH; (c) the selection of tests; and (d) the nature and number of subjects.

Table 3

Contingency Coefficients (C) PH vs. Ratings

Traits	C	P
Active	.605	>.05
Impulsive	.655	>.05
Dominant	.654	>.05
Stable	.676	>.05
Sociable	.585	>.05
Job-effectiveness	.513	>.05
Promotion Possibilities	.697	>.05
Adjustment to Others	.614	>.05

Table 4

Contingency Coefficients (C) Test Results vs. Ratings

Traits	C	P
Active	.753	>.05
Impulsive	.749	>.05
Dominant	.740	>.05
Stable	.733	>.05
Sociable	.746	>.05

The main problem with the criterion has already been mentioned. Regardless of the absence of certain refinements, however, there is no reason to believe that the ratings were any poorer criteria for PH than for the tests.

The method of quantification of PH may be a source of error. However, the selection of the criterion traits was made on a basis which should not only be fair to PH but reasonably important to the selection process. The general agreement among judges indicates that the quantification of PH reports in terms of the selection of traits was adequately accomplished. Surely it would seem reasonable to assume that if a group of experienced psychologists encountered difficulty in making a useful interpretation of PH reports, personnel workers would also have considerable difficulty in making a valid decision on the basis of the information contained in them. Specifically, if the method of quantification were to be suspected as a cause of error, one would *not* expect to find this factor operat-

Table 5

Contingency Coefficients (C) PH vs. Test Results

Traits	C	P
Intelligence		
Speed	.688	>.05
Power	.757	>.05
Active	.685	>.05
Vigorous	.646	>.05
Impulsive	.709	>.05
Dominant <sup>1</sup>	.673	>.05
Stable	.694	>.05
Sociable	.787	<.01
Reflective	.698	>.05
Self-sufficiency	-.702	>.05

<sup>1</sup> From the Thurstone Temperament Schedule.

Table 6

Contingency Coefficients (C) PH vs. Ratings Compared with Test Results vs. Ratings

Traits	PH	Test Results	Differences
Active	.605	.753	.148
Impulsive	.655	.749	.094
Dominant	.654	.740	.086
Stable	.676	.733	.057
Sociable	.585	.746	.161

ing in the case of the comparison between PH estimates of intelligence and the intelligence test results, since the PH estimates were given in terms of Wechsler-Bellevue intelligence quotients. Table 5 indicates no tendency for the relationship between PH and the test results to be higher in the case of intelligence than in that of other traits. In fact, the *only* significant relationship obtained appears on a trait which was quantified by the judges (sociability).

To continue, only test scores already available could be utilized, and certain traits measured were not well adapted to use in a rating scale. The effect, however, would only be to minimize the apparent effectiveness of the objective measures and to make the PH look relatively better.

Finally, the subjects were neither a large nor a random sample from the industrial population. Since they were originally selected because they possessed some outstanding traits or characteristics (an advantage for the PH analyst if he were aware of it), they are by definition a very heterogeneous group. This great variability among them has undoubtedly operated to inflate all the estimates of association herein reported. It thus seems that with comparable subjects and a larger N many of the criterion relationships of Tables III and IV would have been significant. It is difficult, however, to imagine that the apparent merit of PH vs. objective tests was largely influenced by the present choice of subjects.

#### Summary

An investigation was conducted to compare the Worthington Personal History Blank



(PH) and objective test results with respect to their relative efficiency in predicting associates' ratings of 47 employees of an Iowa publishing company. By and large, neither PH nor the objective test results were significantly related to the criterion ratings; however, the coefficients of contingency for the comparison of objective tests with ratings were consistently higher than those obtained for the comparison of the PH with ratings. This difference was significant at the .06 level of confidence.

It was concluded that, under the conditions of the present study, the efficacy of the objective tests employed was at least as great as that of the PH, and very probably greater. It would thus seem that the use of the more

expensive PH is not warranted in terms of cost.

Received May 8, 1953.

### References

1. Moses, L. E. Non-parametric statistics for psychological research. *Psychol. Bull.*, 1952, **49**, 133-136.
2. Schofield, W. Research in clinical psychology: 1951. *J. clin. Psychol.*, 1952, **8**, 255-261.
3. Spencer, G. J. and Worthington, R. E. Validity of a projective technique in predicting sales effectiveness. *Personnel Psychol.*, 1952, **5**, 125-144.
4. Swint, E. R. and Newton, R. A. The personal history,—a second report. Reprinted from *J. Ind. Train.* Jan.-Feb., 1952.
5. Worthington, R. E. *Use of the personal history form as a clinical instrument*. Unpublished Ph.D. dissertation. The University of Chicago, June, 1951.

# A Factor-Analytic Study of Supervisory and Group Behavior<sup>1</sup>

Robert C. Wilson, Wallace S. High, Helen P. Beem

*University of Southern California*

and

Andrew L. Comrey

*University of California, Los Angeles*

A series of studies designed to isolate factors related to organizational effectiveness has been in progress at the University of Southern California (1, 2, 7). The general approach has involved the selection of a number of similar work units or organizations which would be divided into "high," "medium," and "low" groups with respect to criterion data of effectiveness. Questionnaires have been administered to individuals within the work units and the data analyzed against the criterion to determine if the individuals in the more effective work units answer the questions differently from those in the less effective work units.

During the course of this work, the questionnaires employed have gone through several stages of development. The current form involves the use of groups of homogeneous items or "dimensions" developed for the purpose of assessing characteristics of organizations hypothesized to have some relationship to their effective operation.

The dimensions have been revised from one study to the next, usually in the direction of increasing item homogeneity. This process has been carried out on the basis of item analysis information. Intercorrelations among dimension total scores, however, have made it clear that the dimensions overlap considerably, despite differences in the apparent content of the items making up those dimensions.

It was decided that a factor-analytic study of the principal dimensions would provide further information about the relationships

among them and provide a basis for further revision and culling to yield a more economical coverage of the domain. Further, it was believed that such a study would suggest areas which might be explored more fully by developing new dimensions.

## Procedure

*The Sample.* The questionnaire was administered to 98 civilian journeymen at the Long Beach Naval Shipyard. The journeymen are skilled tradesmen who work on all phases of ship overhaul, repair, and construction for the U. S. Navy. Biographical data revealed that medians for the following variables were: (1) age, 40.2; (2) highest school grade reached, 11.6; (3) months worked for present supervisor, 12.6; (4) years in shipyard, 5.0; and (5) years in civil service, 5.8.

*The Questionnaires.* Thirteen dimensions, or homogeneous groups of multiple-choice items, were factor-analyzed. The measuring instrument for each dimension contained six or eight items put in the following objective-item form:

If some worker gets too "eager," employees put pressure on him to make him quit working so hard: 1. always; 2. usually; 3. sometimes; 4. rarely; 5. never.

In this item, as with all others, the five response categories were arranged on a continuum from a response which expressed infrequency or very little of the variable in question, in this case *Lack of Informal Pressures to Restrict Production*, to a response which expressed frequency or a great deal of the variable in question. The responses were arbitrarily weighted from one to five, according to the number preceding the response.

To facilitate computation of reliability estimates, the six or eight items for each dimension were separated into comparable halves or sub-dimensions of 3 or 4 items. This also supplied more variables for the factor analysis.<sup>2</sup> A total

<sup>1</sup> This research was carried out under Contract N6-ONR-23815 between the University of Southern California and the Office of Naval Research. The opinions expressed are our own and are not necessarily shared by the Office of Naval Research. The project is directed by J. M. Pfiffner, with J. P. Guilford and H. J. Locke as associate responsible investigators.

<sup>2</sup> If each dimension does represent a separate factor, an opportunity has thus been provided for the factor to come out. If doublet factors appear for these pairs of sub-dimensions, they can be regarded as specific factors for this analysis.



score for each sub-dimension was obtained by adding the scores assigned to the responses of an individual for the items in that particular sub-dimension.

*The Variables.* The sub-dimensions included in the analysis were composed of items designed to reveal two kinds of information: (1) perceived characteristics of the respondent's supervisor in his relations with employees; and (2) attitudes and interactions among members of the respondent's work group. A sample item has already been given for one pair of sub-dimensions, 11 and 12, *Lack of Informal Pressures to Restrict Production*. For the other sub-dimension pairs, sample items are given below. When reference is made to a third person, e.g., "he" or "him," the person referred to is the respondent's supervisor. (1 and 2) *Pride in Work Group*: You are proud of the work record of your unit: not at all . . . very much; (3 and 4) *Absence of Dissension*: There are people in your unit who refuse to speak to each other: several . . . none; (5 and 6) *Friendly Group Atmosphere*: There is friendly kidding between people in your unit: not much . . . very much; (7 and 8) *Group Cohesion*: People in your unit act as a group to get things they want: never . . . frequently; (9 and 10) *Intensity of Informal Control*: There are certain workers in your unit, besides the boss, who seem to lead the others: not at all . . . very much; (11 and 12) see above; (13 and 14) *Participation*: He is willing to listen to your ideas: never . . . always; (15 and 16) *Lack of Arbitrariness*: He hates to have employees disagree with him: always . . . never; (17) *Non-Apprehension of Authority*:<sup>3</sup> The employees seem to be afraid of him: very much . . . not at all; (18 and 19) *Being Informed*: He passes on interesting bits of information he gets from the front office: never . . . very frequently; (20 and 21) *Feedback*: He lets you know how you are doing: almost never . . . very frequently; (22 and 23) *Attitude Toward Safety Enforcement*: He tries to see that safety rules are observed: almost never . . . very frequently; (24 and 25) *Social Nearness*: He has close friends among his employees: none . . . four or more.

In the above items, the response given a weight of "1" on the dimension appears first, with the response given a weight of "5" following. Intermediate responses have been omitted to save space.

Four of the nine extracted factors gave some appearance of generality as regards the dimensions included in this analysis. Three of the factors were largely specific to a dimension pair and two were residual factors.

<sup>3</sup> One of the *Non-Apprehension of Authority* sub-dimensions was not included in the factor analysis because of lack of item homogeneity.

## Interpretation of the Factors

The factors are presented in the approximate order of their clarity. Interpretations rest upon those variables with loadings of .40 and above. The factor loadings of dimensions defining the factors are given in Table 1.

Factor I. *Supervisor-Subordinate Rapport*. The dimensions with significant loadings on this factor seem to reflect the extent to which a consultative, communicative type of relationship exists between the supervisor and his subordinates. The items on the *Participation* dimension are principally concerned with the supervisor's receptiveness to the ideas and opinions of his subordinates. The *Lack of Arbitrariness* dimension was intended to measure the degree to which the supervisor is not dogmatic and arbitrary regarding his orders and decisions. *Non-Apprehension of Authority* reflects the extent to which subordinates are not afraid of their supervisor. *Feedback* concerns the extent to which the supervisor informs his subordinates as to what he expects of them and lets them know how well they are meeting his expectations. *Being Informed* concerns the extent to which the supervisor tells his subordinates about things that are going on in the organization which may be of interest to them.

This factor seems to reflect the so-called "human relations" approach to supervision. The composition of this factor is quite similar to that of a factor called "Consideration" reported by Fleishman (3) and Gekoski (4). Gekoski attributes the factor to an unpublished factor analysis by Shartle and Hemphill (5). The "Consideration" factor is described as including "behavior which is indicative of friendship, mutual trust, respect, and a certain warmth between the supervisor and group" (4).

An alternative interpretation of the *Supervisor-Subordinate Rapport* factor is that it represents halo effect. Since all the dimensions appearing on the factor involve judgments about the supervisor, there is undoubtedly some tendency for the respondents to rate a supervisor uniformly high or low. On the other hand, substantial correlations among these rating variables might well be expected quite apart from any halo effect, in

that the person who utilizes participation in supervision is less likely to be arbitrary, probably tends to keep his subordinates informed, and so on. Perhaps the most likely interpretation of this factor is that it represents the combined effects of a common factor of *Supervisor-Subordinate Rapport* and halo.

Factor II. *Congenial Work Group*. Factor II represents the degree to which there is absence of discord and the presence of friendly interaction among individuals and groups of individuals within the work unit. The items on *Absence of Dissension* are concerned with the lack of negative interactions, such as friction between workers, bad feeling, refusal to speak to one another, etc., while most of the items on *Friendly Group Atmosphere* emphasize the positive aspect of worker relations, i.e., friendly kidding, good feeling, and high morale in the unit. The majority of items on *Lack of Informal Pressures to Restrict Production* concerns the absence of animosities toward workers who are more productive than the others in the group. If informal work standards have been generated within the group, and social approval is withheld from those who do not conform to these norms, some antagonism or dissension may be present.

Factor III. *Informal Control*. This factor represents the extent to which certain individuals outside the official chain of command exercise influence over the work group, or, more generally, the degree to which the group contains informal status hierarchies. Although the dimension *Intensity of Informal Control* contains items regarding several aspects of the influence of fellow employees upon the individual, the major component appears to be reflected in such items as: "There are certain workers in your unit besides the boss who seem to lead the others." Such informal or indigenous leadership in work groups is a common phenomenon, documented by much research in this area. The loading of *Participation* on this factor suggests that this type of informal leadership is more likely to occur in a group whose supervisor is regarded as receptive to the ideas and suggestions of employees about the work. If the informal leadership is operating in the

work situation, rather than simply in social situations, its emergence in such a situation may be due either to the supervisor's delegation of authority or to his lack of vigorous leadership.

Factor IV. *Group Unity*. Factor IV appears to represent the degree to which the group works together for a common purpose and the extent to which the group is ready to take action as a unit either on behalf of one of its members or on behalf of the group itself. *Group Cohesion* and *Friendly Group Atmosphere* had substantial loadings on this factor for both dimension halves, with the former dimension contributing more heavily. The presence of the latter dimension on this factor suggests that a congenial interaction among group members is a necessary condition for a closely-knit work group. The presence of *Lack of Arbitrariness* possibly indicates that non-dogmatic behavior on the supervisor's part may be conducive to friendly informal organization among the workers.

Factors V, VI, and VII. These factors emerged as doublet factors defined by the three pairs of variables, *Attitude Toward Safety Enforcement* A and B, *Social Nearness* A and B, and *Pride in Work Group* A and B. The fact that they emerged as separate factors is evidence that they are measuring something different from the other variables in the analysis. They will serve the purpose of indicating areas in which we need to construct additional dimensions. In further analyses more general factors may occur in these areas.

#### Summary

Questionnaires were given to 98 skilled tradesmen at a naval shipyard. Items in the questionnaires were grouped to measure: (1) supervisory practices in relations with employees such as *Participation*, *Lack of Arbitrariness*, *Being Informed*, *Feedback*, *Attitude Toward Safety Enforcement*, and *Social Nearness*; and (2) attitudes and interactions of the members of the work group, such as *Pride in Work Group*, *Absence of Dissension*, *Friendly Group Atmosphere*, *Group Cohesion*, *Intensity of Informal Control*, *Lack of Informal Pressures to Restrict Production*, and *Non-Apprehension of Authority*. Each of



these item groups, or dimensions, was divided into two item pools, each pool containing three or four items. Twenty-five of these questionnaire variables were intercorrelated and factor analyzed by the centroid method. Rotation of the centroid axes to meaningful positions was carried out.<sup>4</sup>

The variables calling for evaluation of supervisory practices all emerged on a factor called *Supervisor-Subordinate Rapport* which appeared to reflect a consultative, communicative type of relationship between the supervisor and his subordinates.

Three other important factors were related to relationships among the workers themselves, *Congenial Work Group*, *Informal Leadership*, and *Group Unity*. The first of these was concerned primarily with the degree to which there is informal leadership in the group, and the third reflected the extent

to which the group is unified and is ready to take action as a unit to get something for the group or for one of its members. The remaining identifiable factors were doublets, defined by the paired sub-dimensions, *Attitude Toward Safety Enforcement*, *Social Nearness*, and *Pride in Work Group*.

Received June 8, 1953.

### References

1. Comrey, A. L., Pfiffner, J. M., and Beem, Helen P. Factors influencing organizational effectiveness. I. The U. S. Forest survey. *Personnel Psychol.*, 1952, 5, 307-328.
2. Comrey, A. L., Pfiffner, J. M., and Beem, Helen P. Factors influencing organizational effectiveness. II. The Department of Employment survey. *Personnel Psychol.*, 1953, 6, 65-79.
3. Fleishman, E. A. The description of supervisory behavior. *J. appl. Psychol.*, 1953, 37, 1-6.
4. Gekoski, N. Predicting group productivity. *Personnel Psychol.*, 1952, 5, 281-292.
5. Report Number 6, RF Project 403. Research Foundation, The Ohio State University, 1951.
6. Thurstone, L. L. *Multiple factor analysis*. Chicago: The University of Chicago Press, 1947.
7. Wilson, R. C., Beem, Helen P., and Comrey, A. L. Factors influencing organizational effectiveness. III. A survey of skilled tradesmen. *Personnel Psychol.*, 1953, 6, 313-326.
8. Zimmerman, W. S. A simple graphical method for orthogonal rotation of axes. *Psychometrika*, 1946, 11, 51-55.

<sup>4</sup>To reduce printing costs, the tables of intercorrelations, complete rotated and unrotated factor loadings and the complete set of items (13 pages) have been deposited with the ADI Auxiliary Publications Project. Order Document No. 4116 from Chief, Photoduplication Service, ADI Auxiliary Publications Project, Library of Congress, Washington 25, D. C., remitting \$1.75 for microfilm (images 1 inch high on standard 35 mm. motion picture film) or \$2.50 for photocopies (6 × 8 inches) readable without optical aid. Advance payment is required. Make checks or money orders payable to: Chief, Photoduplication Service, Library of Congress.



## The Check List as a Criterion of Proficiency<sup>1</sup>

Arthur I. Siegel

*Institute for Research in Human Relations, Philadelphia*

The check list represents a particularly attractive tool for measuring a man's ability to perform a task. A performance check list is prepared by analyzing a task into the component actions which a man performs in order to complete the task. In some cases the task involves making something. In this case an end product evolves and the end product is also analyzed in terms of its adherence to certain prescribed standards and its freedom from defect. An examinee receives credit for each of the analytic performance components with which he conforms and each of the elementalistic aspects of his end product that meets prescribed standards. A total task score is derived by adding a man's credits on each of the analytic components of performance and end product adherence to prescribed standards.

For instance, a performance check list for welding consists of items relating to the way the welder performs his job (e.g., "cleans base metal and rods," "adjusts oxyacetylene regulators to 4-5 pounds," "preheats base metal," etc.); the safety precautions the welder follows (e.g., "does not open acetylene cylinder valve more than 1½ turns," "uses goggles when welding," "makes sure fire extinguisher in area before igniting torch"); and items relating to adherence of the final weld to prescribed standards (e.g., "bead width 3-5 times metal thickness," "bead height 25-50% of metal thickness," etc.). The welder is given credit for each of the items with which his performance or final weld conforms, and his total score is the sum of these credits. The problem, as mentioned by Thorndike,<sup>2</sup> is that there may be aspects of a job that are lost in the analytic approach, so that scoring the elementary items does not

give an entirely adequate evaluation. Specifically, the scores obtained by subjects scored in an analytic manner may not correlate highly with the over-all, "clinical" judgments of experts as to the quality of a final product produced.

If performance check list scores do correlate highly with expert, "clinical" judgments or rankings of end products, the performance "check list" is to be preferred. This follows since more objectivity may be introduced into the check list, examiner reliability may be increased by the check list, test reliability may be increased by the check list and less background and less experience in the particular test task is required by the examiner who uses the check list than by the examiner who makes a "clinical" appraisal. Moreover, if performance in process is checked as well as the quality of the final product, certain insights may be gained which could be missed by only an over-all, final appraisal of end products.

### Method

Performance check lists<sup>3</sup> were constructed for four tasks: aluminum butt welding; patching a hole in plastic; splicing a cracked aircraft channel; and aircraft fabric repair. The inter-examiner reliabilities for both the aluminum butt welding and fabrics lists were coincidentally at .92. The inter-examiner reliabilities for the plastics and channel splicing lists were not ascertained. However, the inter-examiner reliabilities on four other check lists similar to the ones herein discussed ranged from .91 to .97. Likewise, the intra-examiner reliability by retest methods for measurements made on the adherence of end products to prescribed standards was ascertained only for the welding and fabrics lists.

<sup>1</sup> The data herein reported are a small portion of the data gathered under Contract Nonr-872(00) between the Institute for Research in Human Relations and the Office of Naval Research. The opinions expressed are those of the author and do not necessarily represent the opinions of the Office of Naval Research or of the Naval Service.

<sup>2</sup> Thorndike, R. *Personnel selection*, New York: John Wiley, 1949.

<sup>3</sup> To save printing costs, the performance check lists as well as the correlational matrices upon which our discussion is based have been deposited with the ADI Auxiliary Publications Project. Order Document No. 4038 from Chief, Photoduplication Service, ADI Auxiliary Publications Project, Library of Congress, Washington, D. C., remitting \$1.75 for 35 mm. microfilm or \$2.50 for 6 by 8 inch photocopies.

These intra-examiner reliabilities were .93 and .87 respectively. Intra-examiner reliability for observations made of performance in process is difficult to obtain by the retest method. This difficulty follows because of the relative impossibility of having an examinee perform the same task in exactly the same manner on two separate occasions. However, by a motion picture technique, the intra-examiner reliability for performance in process was found to be .93 for the welding test. The intra-examiner reliabilities for performance in process of the remaining lists were not determined.

The aluminum butt welding, plastic patching, channel splicing, and fabric repair tasks were first administered to 15 aviation structural mechanics at the Naval Air Technical Training Unit, Memphis. Four of the subjects held the Naval rate of "striker," four held the rate of third class, four the rate of second class and three the rate of first class Aviation Structural Mechanic. All of the jobs represented in these tests are tasks which Naval Aviation Structural Mechanics typically perform. The mean service length of the examinees was 54.6 months with a standard deviation of 38.3 months. The mean scores and standard deviations for the group on each of the tests follow: aluminum butt welding, mean 15.0, sigma 6.9; splicing a cracked channel, mean 44.3, sigma 6.8; plastic repair, mean 20.8, sigma 6.5; and fabric repair, mean 33.4, sigma 9.0. The examiners were Chief Aviation Structural Mechanics who held instructors' billets at the Naval Aviation Structural Mechanics School, Memphis. The examinees were unknown to the examiners prior to the testing situation.

The end products produced by each examinee on each of these tests were taken to the Naval Air Station, Atlantic City. At Atlantic City, five Chief Aviation Structural Mechanics were asked to rank, from best to worst, the end products produced by the examinees on each of the tests.

#### Results

The correlations of the rankings of the Chief Petty Officers who ranked the end products from each test with the rankings produced by our analytic and synthetic ap-

proach and also the correlations between the rankings of the various chiefs who ranked the end products at Atlantic City were calculated. All of these correlations were rank difference correlations. If the experts (Naval Chief Aviation Structural Mechanics) agreed among themselves more than they agreed with the rankings produced by the analytic and synthetic approach, then the analytic and synthetic approach has lost something that these experts considered to be important in making their rankings. On the other hand, if the experts agreed with the rankings produced by the check list as much as they agreed among themselves, then little has been lost by the analytic and synthetic approach.

The median rank difference correlation between the rankings of the chiefs at Atlantic City of the end products from each test was then obtained. The median rank difference correlation of the chiefs' rankings with the rankings produced by our analytic and synthetic check list approach was also calculated. These *rhos* are presented in Table 1.

For the welding test, the median rank difference correlation of the chiefs' rankings with the rankings produced by the analytic and synthetic approach was .41. For the plastics test, the structural maintenance test and the fabrics test, the median rank difference correlations of the chiefs' rankings with the rankings produced by the analytic and synthetic approach were .66, .26 and .33, respectively. For the welding test, the plastics test and the structural maintenance test, the median rank difference correlations between the chiefs' rankings were .95, .89, and .29, respectively, and these three *rhos* were greater

Table 1

Median Correlations Between Chiefs and Between Rankings of Chiefs with Analytic Approach

Test	Median Rho Between Chiefs	Median Rho of Chiefs with Analytic Approach
Welding	.95	.41
Plastics	.89	.66
Structural Maintenance	.37	.26
Fabrics	.29	.33



than the correlation of the chiefs' rankings with the rankings produced by the analytic and synthetic approach. However, the reverse was true for the fabrics test; median  $\rho$  between chiefs .29, median  $\rho$  of chiefs with analytic approach .33. All median rank difference correlations were then converted to product moment correlations and the product moment correlations transformed to  $z$ 's ( $r$  to  $z'$  transformation). The significance of the difference was then calculated between the  $z$ 's which represented the median correlations between the chiefs' rankings of the end products of each test and the  $z$ 's which represented median correlations of the chiefs' rankings with the rankings produced by the analytic and synthetic approach for the same tests.

Of the four tests of significance calculated, only the difference between the median correlation between the rankings of the chiefs and the median correlation of the chiefs' rankings with the rankings produced by the analytic and synthetic approach for the welding test was statistically significant. However, at this juncture, it is well to point out that three of the four median rank difference correlations between the chiefs' rankings were greater than the median correlations of the chiefs' rankings with the rankings produced by the analytic and synthetic approach.

Thus it seems that we were not able to demonstrate conclusively that the Chief Petty Officers in our sample agreed with the rankings produced by the analytic and synthetic approach more than they agreed with each other. On the other hand, in view of the bias (here controlled) that usually enters into judgments of end products, when judgments are made in actual work situations, it seems probable that any loss indicated by the three lower correlations of the chiefs' rankings with the rankings produced by the analytic and synthetic approach as compared with the correlations between the chiefs' rankings would be compensated for if this bias had been allowed to operate here as a confounding variable. Moreover, some procedural elements were not apparent to the chiefs when they made their judgments of the final products. For instance, a man may break safety rules and lose a certain amount of credit. Infrac-

tions such as these are not seen when making appraisals of end products, but may be too important to omit from consideration when estimating a man's real work ability. Since the chiefs who did the ranking had only end products to evaluate, it seems probable that part of the loss indicated by the lower correlations between the chiefs' rankings and the rankings produced by the analytic and synthetic approach as compared with the between-chiefs' correlations may also be assignable to distortions in the rankings of the analytic and synthetic approach due to poor care and use of equipment, violation of safety precautions, etc. on the part of the examinees.

### Summary

As mentioned by Thorndike, there may be aspects of a job that are lost in breaking down a job in terms of the elemental, analytic components comprising the job. If this is so, then scoring the elemental, analytic components of a task does not give an entirely adequate evaluation. Therefore, an investigation was performed into the relationship between total scores assigned via a scoring of the elemental components of a job (performance check list approach) and over-all, "clinical" judgments of experts (Naval Chief Petty Officers) as to the quality of a final product produced. Rank difference correlations were calculated between the scores assigned via a performance check list and judgments of experts as to the quality of final products. Inter-correlations between the rankings of the experts on the quality of end products produced were also calculated. The following conclusions seem warranted.

1. Three out of four median correlations between the rankings of Naval Chief Aviation Structural Mechanics were not significantly different from correlations of the chiefs' rankings with scores obtained by an analytic and synthetic approach.

2. Although no statistical differences were shown in three of the four pairs of rank difference correlations under consideration, the tendency was toward greater agreement between experts' rankings than between the rankings of the experts and the rankings of the analytic and synthetic approach.

Received May 26, 1953.

# Identification and Prediction of Two Training Criterion Factors<sup>1</sup>

Warren R. Graham

New York, N. Y.

The problem of identifying and predicting the variables which are involved in the successful completion of a comprehensive training program was attacked in this study. Twelve examination scores from six courses given as part of the Naval Pre-Flight training program were defined as the training criterion.

The nature of the criterion was studied by the Thurstone centroid factor method. The pre-flight examinations criterion was hypothesized to be composed of two or more factors which could be identified and separately predicted. An attempt was made to predict the factors by an adaptation of the Doolittle multiple correlation method.

## Procedure

In the parent study (3) a battery of standard tests was administered to entering students, and scores from these predictors were correlated with subsequent classroom and final examination grades (criterion variables). The present study employs the ten predictors which had produced regression coefficients indicating that they were the best predictors of the criterion variables.

The sample studied consisted of 399 students in four classes who were homogeneous in having met the following requirements: (1) minimum physical standards; (2) minimum standards on a Naval Aviation selection battery; (3) completion of two years of college; and (4) between 19 and 25 years of age.

The predictor and criterion variables employed were as follows:

### A. The Predictor Variables:

2. *Language Skills Test, G. E. D.*, college level.
6. *Minnesota Clerical—Name checking*, (1946).
7. *Mathematics—Diagnostic*, (Ability to perform simple computations).

8. *Mathematics—Reasoning*, (No high school algebra or geometry required).
9. *Mathematics—Total*, (Combined score for tests 7 and 8).
10. *American Council on Education Psychological Exam., 1947, Q Score*, (Ability to handle quantitative and other nonverbal problems).
11. *Above, L Score*, (Linguistic or verbal abilities).
12. *Above—Total*, (Combined score for tests 10 and 11—general academic ability).
13. *Aviation Classification Test*, (General academic ability).
14. *Mechanical Classification Test*, (Ability to solve mechanical problems).

### B. The Criterion Variables:

16. (Daily),<sup>2</sup> 17. (Exam.),<sup>3</sup> *Aerology*, (Meteorology).
23. (Daily), 24. (Exam.), *Engines*, (Parts, functions, instrument checks, weights and balances, etc.).
26. (Daily), 27. (Exam.), *Essentials of Naval Service*, (Regulations, military courtesy, leadership, etc.).
30. (Daily), 31. (Exam.), *Principles of Flight*, (Aircraft nomenclature, lift, drag, stability, compressibility, etc.).
33. (Daily), 34. (Exam.), *Navigation, Dead Reckoning*, (Earth's coordinates, variation and deviation, winds, fixes, etc.).
35. (Daily), 36. (Exam.), *Navigation, Celestial*, (Astronomy, celestial triangle, latitude by polarities, sextant, astro-compass, time and radio signals, etc.).
39. *Navy Grade*, (a combined score for all criterion variables).

**Statistical Technique.** The twelve criterion examinations were factored, yielding three factors, of which two could be rotated to psychologically meaningful patterns of factor loadings, using Thurstone's centroid method and graphic rotation.

The intercorrelations of the ten predictor variables which had been shown to be most related to the twelve criterion examinations in the parent study (3) were then added to the criterion intercorrelation matrix. This produced the rectangular intercorrelation matrix shown in Table 1. The predictors were related to the criterion factors in such a way as to obtain rotated predictor factor loadings without changing the criterion

<sup>1</sup> The author is indebted to Dr. Harold A. Edgerton for guidance and suggestions offered during analysis of the data. The materials treated herein were collected by Richardson, Bellows, Henry and Co., Inc., under Office of Naval Research Contract N 7 onr-383, TO-I, with the cooperation of the Special Devices Center and the Pensacola Naval Air Training Command. Interpretations and opinions are the author's and should not be construed as having U. S. Navy endorsement.

<sup>2</sup> *Daily Grades.* Averages of grades from instructor-prepared weekly quizzes.  
<sup>3</sup> *Exam. Grades.* Grades on a two-hour final exam, prepared by a central examining board.



sults obtained for the separate analysis of the criterion examinations.\*

When the factor loadings for the predictors had been obtained, they were considered to be validity coefficients between the predictors and the criterion factors. The Doolittle multiple correlation method was then employed to get a prediction of each of the two criteria obtained (the criterion factors) by each predictor in terms of standard partial regression coefficients (beta weights). The regression coefficients represent the degree to which a predictor will predict each criterion when the influence of the other predictors has been partialled out or controlled.

Thus, instead of predicting each of the twelve criterion variables or a composite of them, two factors which economically and accurately represent them have been predicted.

The composite variables (Math-Total, Navy Grade and American Council on Education—Total) for which factor loadings were computed, were not included in the Doolittle solution to avoid contamination due to spurious correlations with the tests which compose them.

No measures of the reliability of the criterion variables are available.

### Results

Tables 2 and 3 present the factors which resulted from analysis of the criterion examinations. Two of the three factors are psychologically meaningful when rotated to simple structure. These are:

Factor I. *Navigation*. Table 3 indicates that the highest factor loadings occur for the four navigation criterion examinations (variables 33, 34, 35, 36). As shown in Table 4, this factor was best predicted by the Minnesota Clerical Test-Name Checking, var. 6; the Mathematics Diagnostic Test, var. 7; and the American Council on Education Test-Q Score, var. 10. It was negatively predicted by the ACE Test-L Score, var. 11, and the Mechanical Classification Test, var. 14.

In general, the Navigation factor is best predicted by the tests of arithmetic and clerical (name-checking) abilities, and negatively predicted by the tests of linguistic abilities.

Factor II. *Verbal Reasoning*. Table 3 shows high loadings for all variables except those of an obviously non-problem solving nature. This indicates that general ability is of importance to success in courses in the Pre-Flight curriculum. Table 4 indicates that this factor is best predicted by the

\* See the Technical Appendix below for a description of this technique.

Table 2  
Centroid Factor Matrix (before rotation)

Predictors	I	II	III
2. Lang. Skills, GED	287	074	-046
6. Minn. Clerical Test, Names	257	-144	-080
7. Math. Diagnostic	346	-280	007
8. Math. Reasoning	542	-105	-073
9. Math. Total	511	-220	-225
10. Am. Counc. Ed.—Q Score	318	-191	-034
11. Am. Counc. Ed.—L Score	311	240	-003
12. Am. Counc. Ed.—Total	362	057	-184
13. Aviation Class. Test	441	023	-049
14. Mech. Class. Test	361	104	281
Criterion			
16. Aerology Daily	673	135	-145
17. Aerology Exam.	648	130	078
23. Engines Daily	701	159	148
24. Engines Exam.	652	161	101
26. Ess. Naval Serv. Daily	597	145	-243
27. Ess. Naval Serv. Exam.	444	158	-247
30. Prin. Flight Daily	636	306	206
31. Prin. Flight Exam.	590	292	125
33. Nav. Dead Rec. Daily	744	-317	156
34. Nav. Dead Rec. Exam.	462	-435	-148
35. Nav. Celest. Daily	630	-372	120
36. Nav. Celest. Exam.	543	-419	-093
39. Navy Grade	936	-090	-070

Mathematics Reasoning Test, var. 8, the Aviation Classification Test, var. 13, and the Mechanical Classification Test, var. 14. The ACE Test-L Score, var. 11, is a less significant predictor than any of these three tests, but it is considerably more effective than the arithmetic and clerical (name-checking) tests. In general, Factor II is best predicted by tests calling for reasoning ability.

The Language Skills Test, GED-College Level, does not contribute to the prediction of either factor.

Factor III. *Unidentifiable Variance*. Although this variable is not clearly interpretable, there is a possibility that it reflects relatively specific variance in the examinations for Essentials of Naval Service, var. 26, 27. It might represent a personality or social factor which has not been sufficiently determined to be identified.

It is concluded that successful completion of the Navy Pre-Flight curriculum is dependent, in part, upon ability to perform simple arithmetic computations accurately, such ability being required to pass courses in navigation. That some clerical ability is required is indicated by the fact that the Minnesota Clerical Test-Name Checking, var. 6, is a

Table 3  
Centroid Factor Matrix (after two rotations)

Predictors	I	II	III
2. Lang. Skills, GED	.005	.295	-.050
6. Minn. Clerical Test, Names	.215	.215	-.035
7. Math. Diagnostic	.325	.280	.090
8. Math. Reasoning	.225	.505	.020
9. Math. Total	.370	.450	-.140
10. Am. Counc. Ed.—Q Score	.250	.275	.025
11. Am. Counc. Ed.—L Score	-.165	.350	-.045
12. Am. Counc. Ed.—Total	.070	.365	-.175
13. Aviation Class. Test	.080	.440	-.035
14. Mech. Class. Test	-.070	.375	.275
Criterion			
16. Aerology Daily	.045	.680	-.140
17. Aerology Exam.	-.010	.660	.072
23. Engines Daily	-.040	.719	.140
24. Engines Exam.	-.045	.675	.100
26. Ess. Naval Serv. Daily	.040	.610	-.245
27. Ess. Naval Serv. Exam.	.000	.470	-.255
30. Prin. Flight Daily	-.205	.685	.165
31. Prin. Flight Exam.	-.180	.640	.090
33. Nav. Dead Rec. Daily	.415	.665	.259
34. Nav. Dead Rec. Exam.	.540	.362	-.023
35. Nav. Celest. Daily	.455	.542	.230
36. Nav. Celest. Exam.	.525	.450	.025
39. Navy Grade	.295	.895	.000

predictor of the Navigation factor. Verbal facility and ability to solve verbal problems are prerequisite to success in all academic courses (including navigation) in the Pre-Flight curriculum.

### Technical Appendix

A correlation matrix of the criterion variables was prepared (Table 1—Criterion Matrix) and subjected to multiple factor analysis by Thurstone's centroid method.

McNemar's (1) criterion for the number of factors to be extracted (that the SD of the residuals should be down to or below the SD of the zero-order  $r$ 's) is difficult to interpret for this study, but it suggests that three factors, and possibly more, should be extracted. The difficulty of interpretation arises in determining how

Table 4  
Regression Coefficients for the Prediction of  
Criterion Factors I and II

	I	II
2. Language Skills, GED	-.016	.022
6. Minn. Clerical Test, Names	.234	.042
7. Math. Diagnostic	.263	.038
8. Math. Reasoning	.036	.297
10. Am. Counc. Ed.—Q Score	.230	-.025
11. Am. Counc. Ed.—L Score	-.367	.123
13. Aviation Class. Test	.075	.205
14. Mech. Class. Test	-.111	.207

low the SD of the residuals must be before it may be considered to meet the criterion of being "down to" the value for the zero-order  $r$ 's. The criterion used in this study was to stop extracting factors when the residuals closely approximated zero.

The zero-order  $r$ 's for the predictors were simply added to the end of the square criterion matrix to form a rectangular matrix (Table 1). The predictor columns were summed and divided by the square root of the sum of the sums of the criterion columns for each factor, thus placing the predictors in the same space as the criterion variables. The predictor residuals were computed and reflected according to the reflections determined by the criterion variables. The predictor factor loadings extracted (Table 2) were then rotated to the identical planes used for the criterion factors.

The Doolittle method was used to obtain standard partial regression coefficients to express the relationships between the predictors and the criterion factors. The factor loadings of each predictor for each factor were placed in the criterion columns (i.e., two criterion columns [one for each factor] were used instead of the usual one, and the predictor factor loadings were treated as validity coefficients). The basic equations were solved for the regression coefficients to estimate the relationship of each predictor variable to each factor.

The principal advantage to be derived from the initial factor analysis of the criterion variables is that the obtained criterion factors may be isolated free from the influence of the variance of the predictor variables. The rotated factor loadings of the predictors indicate which ones are most likely to be related to the criterion factors. Thus, computation of regression coefficients for irrelevant predictor variables may be avoided.

Extension of the Doolittle multiple correlation table to permit the simultaneous computation of regression coefficients for the prediction of several criteria does not alter the results ordinarily obtained for any single criterion. The nature of the computations required permit the prediction of any number of criteria once the computations for the zero-order inter-correlations of the predictors are completed.

Received June 5, 1953.

### References

1. McNemar, Q. On the number of factors. *Psychometrika*, 1942, 7, 9-18.
2. Peters, C. C. and Van Voorhis, W. R. *Statistical procedures and their mathematical bases*. New York: McGraw-Hill, 1940.
3. Edgerton, H. A. *A study of individual differences among naval aviation students*. New York: Richardson, Bellows, Henry and Co., Inc., 1949.
4. Thurstone, L. L. *Multiple factor analysis*. Chicago: University of Chicago Press, 1947.



## Rater and Technique Contamination in Criterion Ratings

Gloria H. Falk and A. G. Bayroff<sup>1</sup>

*Personnel Research Branch, The Adjutant General's Office, Department of the Army,  
Washington, D. C.*

An important consideration in validation studies is the degree to which the procedures for obtaining the criterion measures are independent of the predictors. Should predictor test scores enter into the determination of criterion scores then the correlation between predictor scores and criterion scores will be artificially increased. Similarly, if the criterion measure is a rating, the validity will be inflated if the raters base their evaluations on prior knowledge of the predictor scores.

When both the predictor and criterion measures are ratings, the problem of criterion contamination may be critical. It may take the form of rater contamination as, for example, when both the predictor raters and the criterion raters are the same persons. Or, criterion contamination may take the form of technique contamination. This form of contamination may exist when both the predictor rating and the criterion rating employ the same rating technique. Thus, a graphic predictor rating may correlate more highly with a graphic criterion rating than would a forced-choice predictor rating with a graphic criterion rating. Since the usual criterion rating employs some form of graphic rating, the possibility exists that the graphic rating technique appears more valid than others primarily because both predictors and criteria employ the same measuring technique.

### Problem

The subject of study described here was the comparative influences of two potential

sources of criterion contamination,—similarity of raters and similarity of techniques. An attempt was made to estimate the relative amounts of agreement between two sets of ratings when the two sets involved: (a) the same raters; and (b) the same techniques.

### Method

*Population.* The population consisted of 400 officers (primarily majors and lieutenant colonels) enrolled as students at the Army Command and General Staff College. The objective of this college is to train potential division commanders and general staff officers, and its students represent a highly selected group. The course was 42 weeks long during which the students were in close contact with one another. Each officer served as both rater and ratee.

*Design of the Study.* This study is one part of a larger research program on rating methodology. This report will be limited to those aspects pertinent to the present study. The study permitted a comparison to be made of the amount of agreement between ratings made with identical techniques and those made with different techniques under two general conditions: (a) when the same raters made the two sets of ratings; and (b) when different raters made the two sets of ratings.

*Instruments.* The rating techniques employed in this study were an eight-point graphic scale of over-all value to the service and two versions of the forced-choice technique.<sup>2</sup> The graphic scale was provided with descriptions for each of its eight-points, ranging from "The most outstanding officer I know" as point 1, to "An officer who does not have the calibre that one should reason-

<sup>1</sup> The opinions expressed in this article are those of the authors and do not necessarily express the official views of the Department of the Army. Acknowledgment is made of the participation of various staff members of the Personnel Research Branch, particularly Edward A. Rundquist and Helen R. Haggerty. Acknowledgment is also made of the generous assistance of the Commandant, Command and General Staff College, Fort Leavenworth, Kansas, and his staff, and the officer students in attendance during the gathering of the field data basic to the study.

<sup>2</sup> Schneider, Dorothy E. and Bayroff, A. G. The relationship between rater characteristics and validity of ratings. *J. appl. Psychol.*, 1953, 37, 278-280.

ably expect in an officer" as point 8. To counteract the reluctance of most raters to use the low end of the scale, two of the points below mid-scale value were favorably defined, e.g., "An acceptable officer whose value is limited in some respects." The fact that the entire scale was used may, in part, be attributed to this device.

Both versions of the forced-choice technique had identical items. In one form, the controlled check list (CCL), the items were arranged in two sets of 24 phrases each. The rater selected the 12 most descriptive phrases in each set. In the second form, the forced-choice pairs (FCP), the items were arranged in 24 pairs. Phrases in each pair had similar preference values, but different discrimination values. The raters selected one phrase in each pair.

*Procedure.* The classes consisted of 33-40 officers each. Each officer was assigned 20 other officers of his class to rate on the graphic scale. These assignments were made according to a procedure which approximated random selection, and the order in which the ratings were to be made was specified.

For purposes connected with other studies in this series and not relevant to this study, each class was randomly divided into two groups of raters. One group signed its ratings and was informed that these ratings would be available for official use; the other group did not sign its ratings and was told that these ratings would not be available for official use. Each ratee received half his ratings from one group and half from the other. The results in this study will be presented separately for the two groups as a partial replication device.

Eight days after the graphic ratings were made, each rater re-rated two of his fellow-officers, first on one of the two versions of the forced-choice technique and then on the same graphic scale used a week earlier.

*Different Raters, Same Technique.* Average of the intercorrelation coefficients among graphic ratings from four different raters per ratee was determined.

*Different Raters, Different Techniques.* Average of the correlation coefficients between the forced-choice ratings on the ninth day

Table 1  
Product-Moment Correlations Between  
Sets of Ratings

Raters	Rating Techniques		
	Same	Different	
		Graphic vs.	
	Graphic vs. Graphic	Controlled Check List	Forced- choice Pairs
Same	.82 .69	— .52	.57 —
Different	.30 .25	.29 .24	.23 .26

and graphic ratings from three different raters per ratee on the first day was computed. The forced-choice ratings selected for this analysis were those made by the raters whose graphic ratings were omitted here.

### Results and Conclusions

As shown in Table 1, the highest correlations were obtained for the sets of ratings made by the same raters using the same techniques ( $r = .82, .69$ ). Somewhat lower correlations were obtained for the ratings made by the same raters using different techniques ( $r = .57, .52$ ). The lowest correlations were obtained for the sets of ratings made on the same ratees by different raters using the same techniques ( $r = .30, .25$ ) and by different raters using different techniques ( $r = .29, .26, .24, .23$ ).

It was to be expected, of course, that evaluations of the same ratees by the same raters rendered 8 days apart would be in substantial agreement. It was also to be expected that agreement would be less when the evaluations were made by different raters on the two occasions. However, the significant facts to be noted are these: (a) when the same raters were involved, agreement was greater when the same technique was employed than when different techniques were employed; and (b) when different raters were involved it made no difference whether the same or different techniques were employed.

It appears, therefore, that contamination



in this study was linked to the raters. Contamination resulting from similarity of technique appeared only when the raters were identical and was virtually absent when the raters were different.

In evaluating these findings, the following limitations of the study should be borne in mind: (a) the design of the study did not permit the use of all combinations of raters and technique; (b) it did not permit varying degrees of overlap in the two sets of raters or the time interval between ratings; (c) only a limited variety of techniques were studied; (d) the rater population may not have been

typical of Army raters. Nevertheless the findings were internally consistent and the following generalization stated in terms of criterion contamination may be offered: in validating rating instruments against criterion ratings, rater contamination is more serious than is technique contamination. If the raters who provide the predictor ratings are different from those who provide the criterion ratings, no technique contamination will result. If, however, the raters are the same, technique contamination may appear if the same techniques are used.

*Received April 27, 1953.*

## Validity versus Reliability

Edward K. Strong, Jr.

*Stanford University*

Which is preferable—a test with higher validity and lower reliability or a test with lower validity and higher reliability?

Two recent investigations have raised this issue. Neither is sure but both incline toward giving greater weight to validity than reliability. Clark (1) says, "The evidence thus far presented makes any decision about types of keys to be used rather difficult, since improved separation of groups must be weighed against decreased test retest reliabilities. The alternative conclusion, that low reliability has little meaning in a situation where high validity is obtained, is of course a possibility. Had the estimate of reliability been other than a test retest measure, this would have been an attractive alternative."

Strong and Tucker (4) report: "BS scales have been selected over BO scales on the basis of greater validity. It is believed that validity is more important than reliability, that validity automatically necessitates reliability, and that the measures of internal consistency reported herein are not complete measures of reliability."

Both of these investigations have been concerned with the selection of items for a key to differentiate two groups on the basis of their interests. Clark's aim was to select items, so that different aspects of differences in interests would be represented by the same number of items in each case. He does not claim to have achieved this desideratum nor to have

developed a satisfactory method of doing so, but that such was his aim and some real progress was achieved.

Strong and Tucker developed keys to differentiate medical specialists from medical men-in-general. They found that the original internist medical specialist key did not differentiate internists from psychiatrists to any marked degree. Items were then selected that would differentiate not only internists from medical men-in-general, but also internists from surgeons, pathologists and psychiatrists. The original scales and the revised scales were designated, respectively, BO and BS scales.

A small sample from each of these investigations is given here to illustrate the problem. Table 1 indicates that the iterative and Gulliksen types of key are superior to the original, i.e., customary types of key, as far as validity is concerned but have appreciably lower reliability. Table 2 illustrates the same situation from data of Strong and Tucker. Differences were not great for three pairs of scales but were in the indicated direction. The really serious problem concerned the BO and BS internist scales. Here the BO scale of 278 items had a reliability of .86 in comparison with the BS scale of 69 items and reliability of .69. But the BO scale had a validity of 69 per cent overlap, biserial  $r$  of .47 in comparison with the BS scale which had a 51 per cent overlap and biserial  $r$  of .68.

It is to be noted that in both investigations

Table 1

Separation of Aviation Machinists Mates from Navy Men-in-General (Clark)

Type of Key	Number of Items	Per Cent Overlap		Test-Retest Reliability
		Original	Cross-Validation	
Original	83	65	58	.85
Iterative	42	51	51	.74
Gulliksen	49	56	51	.75



Table 2

Validity and Reliability of BO and BS Scales (Strong and Tucker)

Type of Scale	Average Number of Items	Validity		Reliability (odd-even)
		Per Cent of Total Overlap	Biserial $r$	
4 BO Scales	285	52	.67	.85
4 BS Scales	161	42	.77	.79

a decrease in number of items resulted in decrease in reliability, although the retained items had individually greater reliability and distinctly greater validity.

The data in Table 3 have just been tabulated. Their import raises anew the question whether increased validity can offset decreased reliability. The odd-even reliabilities

Table 3

Odd-Even Reliability of Scales and Test-Retest Correlations over an Average of 18 Years

Scales	Odd-Even Reliability	Test-Retest Correlation N = 663
Engineer	.94	.79*
Life Insurance	.93	.75
Chemist	.91	.79
Sales Manager	.90	.68
Real Estate	.90	.69
Doctor	.87	.76
Farmer	.88	.67
Lawyer	.88	.73
Office	.88	.65
Production Manager	.85	.67
Accountant	.84	.65
Banker	.83	.72
President	.82	.50
Personnel	.82	.54
Public Administrator	.76	.48**

\* Previously reported as .76, based on 203 cases (3).

\*\* N = 248.

of 15 scales for the Vocational Interest Blank (2, p. 78) are given in the table and the corresponding test retest correlations for 663 former college students retested on the average 18 years later. The eight scales with high reliability (.88 to .94) have an average test retest correlation of .73 in contrast to the correlation of .60 for the seven scales with poorer reliability. The rank order correlations between odd-even reliability and test retest correlation is .83.

The conclusion from these data is that if one wishes to have test scores as a basis to predict behavior in the distant future he wants tests that will give as great agreement as possible between scores today and scores in the future and that the reliability of the scale is important in this connection.

Received May 19, 1953.

### References

1. Clark, K. E. *Research on scoring methods for the U. S. Navy Vocational Interest Inventory*. Technical Report No. 5, 1952, Department of Psychology, University of Minnesota.
2. Strong, E. K., Jr. *Vocational interests of men and women*. Stanford: Stanford University Press, 1943.
3. Strong, E. K., Jr. Nineteen-year follow-up of engineer interests. *J. appl. Psychol.*, 1952, 36, 65-74.
4. Strong, E. K., Jr. and Tucker, A. C. The use of vocational interest scales in planning a medical career. *Psychol. Monogr.*, 1952, 66, No. 9.

## Sampling Problems in Studies of Writing Style

Richard D. Powers

*Department of Agricultural Journalism, University of Wisconsin*

The past few years have seen a growing number of studies in which the constituent elements of writing style are examined individually and statistically. The work of Rudolph Flesch (5) and of Edgar Dale (2) in "readability scoring" is perhaps best known, but this paper is primarily concerned with lesser known developments (8, 10) and with the future of stylistic measuring devices. Formulas for readability are only a secondary consideration.

With most style studies has come the necessity for sampling. Application of stylistic analysis to the total content of a book or newspaper may be unnecessary, and in many cases is usually physically impossible. Samples are the alternative.

Sampling theory gives a basis for estimating the minimum size of sample needed for varying degrees of precision. However, the most useful tests can be applied only to *random* samples; in other words, to samples where each unit of the kind being studied had an equal chance of being drawn in the sample.

This brings us to a special problem of sampling in writing style. Usually two kinds of units are involved in an analysis of style:

1. Sentence characteristics—length, form, structure.
2. Word characteristics, and principally the relative difficulty of various words.

Since, in the most popular style-measuring devices, sampling concerns both of these units, almost all sampling has been drawn by drawing a sample of *sentences* and then analyzing these sentences and the words in them as the particular formula requires. If, in future studies concerning only word measurements, the assumption is made that this procedure gives a random sample of *words*, the results could be misleading, depending upon the type of stylistic measurement.

It is apparent that a 100-word sample drawn by using all the words in five randomly selected sentences does not fulfill the conditions of random selection of 100 words. And Baker (1) has presented evidence that selection of 100 consecutive words in paragraphs could bias results of certain kinds of studies. Sampling by sentences in *studies of word characteristics* is actually a type of cluster sampling. It involves a selection of several connected units of analysis at one drawing. As such, it is a restriction of randomness which could cause logical and statistical difficulties in this type of study.

For such a clustered sample, it is an error to apply the formula  $SE = \frac{\sigma}{\sqrt{n}}$  for deter-

mining the precision of the sample. Of course, there are ways to evaluate the precision of a clustered sample. But to do that, we must have additional information, such as the amount and direction of intercorrelations between the elements within the clusters. Data on the intercorrelations between words in sentences are necessarily very meager, so we have no way of knowing how we affect the precision of our samples by clustering.

Some relationships between words in a sentence are obvious, though. For instance, when "the" appears in a sentence, it is usually followed by an adjective or a noun, sometimes by an adverb, but almost never by a verb. And, up to a point, the more words preceding a given word in a sentence, the more the nature of that word is predetermined.

Because of this verbal contextual effect—one word increasing or decreasing the probability that the words which follow will be certain words or types of words—it would seem that some kinds of words might be over-represented when samples of sentences are drawn for studies of word characteristics. Likewise, it would seem that some kinds of words might be under-represented.



Simple random sampling could safeguard against these difficulties. Theoretically, moreover, a study of sentence length must be done by drawing a random sample of sentences; a study of word lengths or proportions of parts of speech calls for a random sample of words; and, theoretically again, a study of clause structure must be made on a random sample of clauses. But we compromise between pure theory and practical considerations almost every day, with no drastic results and often with considerable economy.

### Procedure

This study, then was an empirical attempt to see how "cluster" sampling (sampling by sentences) affected certain arbitrarily selected word variables: (a) representation of different parts of speech in the sample; (b) proportion of "hard" words (defined as words not in Edgar Dale's list of 3,000 words); (c) proportions of words of different syllable lengths (words of one or two syllables were called "short" words for the purposes of this study); and (d) proportions of "structural" words (defined as prepositions, conjunctions, and articles—an exclusive, though not inclusive category).

The two samples were drawn by equally random methods. The sample of 1,000 words was picked one word at a time by a table of random numbers. The sample of 64 sentences was also picked by a table of random numbers (997 words).

For the word sample, the first four numbers of the random number table designated the page number, the next number designated the paragraph number on that page or the following one, and the next two numbers indicated the word in the selected paragraph (or the following one) which was to be drawn into the sample. The sentence sample was selected using the first four random numbers for the page number and the next two numbers for the sentence number on that page or the following one. The samples were drawn from a three-volume report the U. S. Department of Agriculture prepared for a congressional committee (12).

This particular report was used for sampling because the first aim of the study had

been to establish grammatical and vocabulary differences between a "popularized" version and a rather technical version of the same material. Both kinds of writing were contained in the report. The difficulty in determining sample size led to the present study, the original intent having been dropped.

A sample of 1,000 words was the size selected because, on assumptions of simple random sampling, this size assures accuracy in word studies to within three per cent of the sample value 95 per cent of the time when allowing for maximum variability. Most of the time, the precision of such a sample would approach two per cent for measurements of parts of speech. A rough analysis of the accumulated percentages by 50 word subsamples shows that, in general, the sentence sample was more stable than the word sample. That is, the curve levelled off at an earlier point than for the word sample.

The significance of the difference between the two samples for each measure was established by a "t" score—the difference between the two, divided by the standard error of the difference.

### Results

Table 1 shows that for measurement of the proportion of different parts of speech drawing the sample by sentences didn't significantly affect the results. (A word was classified as a particular part of speech on the basis of traditional grammatical rules, with the exception that nouns modifying other nouns modifying other nouns were classified as adjectives.)

However, as shown in Table 2, the results were significantly different for the measurements of proportions of "short" words, "easy" words, and "structural" words. These categories are more rigidly defined and logical reasoning would tell us are more meaningful in measuring such style aspects as the reading ease of writing. No significant differences were obtained when the proportion of nouns, adjectives, and verbs were lumped together in a manner similar to that of forming the "structural words" category. Since the structural words are generally short words that are included in the Dale 3,000 word list,

Table 1

Proportions of the Various Parts of Speech in a Sample Drawn by Selecting all Words from Randomly-drawn Sentences

	Sample Drawn by Words		Sample Drawn by Sentences		Difference		t* Score
	%	N	%	N	%	N	
Nouns	31	313	31	303	—	10	—
Adjectives	20	202	17	166	3	36	1.73
Verbs	14	136	14	137	—	1	—
Adverbs	1	11	1	13	—	2	—
Prepositions	18	185	19	194	1	9	.57
Articles	7	68	8	82	1	14	.94
Conjunctions	5	47	6	56	1	9	.95
Pronouns	4	36	5	46	1	10	1.08

\* Score of 1.96 statistically significant.

these latter three factors are probably inter-related.

Two hypotheses suggested by this part of the study are that the measure of average syllable length per word is a less sensitive gauge of word difficulty than the measure of proportions of words of different syllable lengths, or that it may require huge samples to reveal smaller shades of difficulty. This might suggest that Gunning's *Fog Index* (7) is a readability measure based at least in part on sounder premises than Flesch's readability formula, or that the Flesch formula may require larger samples than we have thought.

It might also merit study in light of the recent controversy between Farr, Jenkins, and Paterson (3, 4) vs. Klare (9) and Flesch (6) as to whether or not a count of one syllable words would suffice to indicate vocabulary

difficulty. Though the actual behavior and evaluation of these two measurements is a subject for further detailed study, Table 3 shows the proportions of words of various syllable lengths obtained in this study:

#### Discussion

In applying the findings of this study to the field of style analysis, it can safely supply only a caution: that as measurements of style variables become more refined, the sampling methods may also have to become more refined. In other words, as style analysis goes from rather crude and vaguely-defined measurements such as proportions of parts of speech proposed by Stormzand and O'Shea (11) to the "psychogrammatical" categories proposed by Sanford (10) or to other refined ratios and categories (8), sampling by the

Table 2

Proportions of "Short" Words, "Hard" Words, and "Structural" Words in a Sample Drawn Randomly by Words and a Sample Drawn by Selecting All Words in Randomly-drawn Sentences

	Sample Drawn by Words		Sample Drawn by Sentences		Difference		t* Score
	%	N	%	N	%	N	
Short Words	73	726	77	768	4	42	2.06
Hard Words	39	386	32	318	7	68	3.29
Structural Words	30	300	33	332	3	32	2.03
Syll. per Word	1.89		1.84		.05		.03

\* Score of 1.96 statistically significant.



Table 3

Proportions of Words of Various Syllable Lengths in a Sample Drawn Randomly by Words and a Sample Drawn by Selecting All Words from Randomly-drawn Sentences

	"Short" Words				"Long" Words					
	1 syll.		2 syll.		3 syll.		4 syll.		5 syll.	
	N	%	N	%	N	%	N	%	N	%
Word Sample	509	51	217	22	143	14	99	10	29	3
Sentence Sample	545	55	223	22	125	13	67	7	32	3

traditional method of drawing words in sentences and in paragraphs should be examined carefully to see that the clustering of units does not adversely affect the results to such a degree that the efficiency of sampling was a false economy.

### Summary

As our techniques of studying language become more refined, we need to take a closer look at our sampling methods.

The samples usually drawn for language studies are made up of clusters of words, in sentences or paragraphs. In some studies, the words so collected have been subjected to further analysis.

Such a sample is a "clustered" sample, and a network of unknown intercorrelations between the words interferes with the known probability that any unit of the universe will be drawn in the sample. Thus, for some word characteristics at least, sampling by sentences would bias the sample of words in an undetermined direction.

Random sampling of words is suggested as a way to sidestep such difficulties in word studies. A comparison of the two sampling methods (clustered and simple random) indicates that a clustered sample significantly overestimated the percentage of "short" words, "structural" words, and "easy" words. It is suggested that the structure of the sentence (the need to have many short and easy

connective words) has imposed an orderliness that has biased the clustered sample.

Received May 29, 1953.

### References

1. Baker, S. J. A linguistic law of constancy. *II. J. gen. Psychol.*, 1951, 44, 113-120.
2. Dale, E. and Chall, J. S. A formula for predicting readability. *Ed. Res. Bull. Ohio St. Univ.*, 1948, 27, 37-54.
3. Farr, J. N., Jenkins, J. J. and Paterson, D. G. Simplification of the Flesch reading ease formula. *J. appl. Psychol.*, 1951, 35, 333-337.
4. Farr, J. N., Jenkins, J. J., Paterson, D. G. and England, G. W. Reply to Klare and Flesch on "Simplification of reading ease formula." *J. appl. Psychol.*, 1952, 36, 55-57.
5. Flesch, R. A new readability yardstick. *J. appl. Psychol.*, 1948, 32, 221-223.
6. Flesch, R. Reply to "Simplification of Flesch reading ease formula." *J. appl. Psychol.*, 1952, 36, 54-55.
7. Gunning, R. D. *The technique of clear writing*. New York: McGraw-Hill Co., 1952.
8. Johnson, W. *People in quandaries*. New York: Harper, 1946.
9. Klare, G. R. A note on "Simplification of the Flesch reading ease formula." *J. appl. Psychol.*, 1952, 36, 53.
10. Sanford, F. H. Speech and personality. *Psychol. Bull.*, 1942, 39, 811-845.
11. Stormzand, M. J. and O'Shea, M. *How much English grammar?* Baltimore: Warwick and York, 1924.
12. U. S. Government Printing Office Report of Research and Related Activities of the U. S. Department of Agriculture, Washington, D. C., 1951.

# Differential Prediction of Academic Success at Brigham Young University<sup>1</sup>

Joics B. Stone

Brigham Young University

The research on prognosis of college academic success (1, 2, 3, 4) has focused on three phases of the problem: (a) prediction of general scholarship; (b) prediction of scholarship in specific subjects or subject groups; and (c) differential prediction in major areas or curricula. The most effective predictor variables have proved to be high school grade-point average, some measure of scholastic aptitude, and an objective measure of high school achievement. Multiple correlations have proved more efficient, generally, than zero-order correlations.

The present study represents an attempt to provide multiple regression equations which can be used in the differential prediction of academic success in four college curricula at Brigham Young University: (a) commerce; (b) elementary education; (c) physical sciences; and (d) social sciences.

## Plan of the Study

**Curriculum Components.** The four curricula studied included the following academic departments:

1. Commerce: accounting, business administration, finance and banking, and marketing.
2. Elementary education.
3. Physical sciences: chemistry, geology, mathematics, and physics.
4. Social sciences: history, political science, and sociology.

**Criterion.** The criterion was selected to conform to those curricula. The curriculum grade-point average (CGPA) was selected as the measure of the criterion. Only courses essential to each curriculum were used in computing the CGPA. A minimum of thirty curriculum credit-hours was required for each student. This minimum constituted from one-half to two-thirds of the departmental major requirement for graduation. The reliability of the criterion was checked

by correlating the CGPA of the first 10 hours of curriculum credit against the total CGPA. For the respective criteria the reliability coefficients were: commerce, .78; elementary education, .82; physical sciences, .79; and social sciences, .68.

**Students.** The commerce curriculum included 102 students; 123 were in elementary education; 133 in the physical sciences; and 78 in the social sciences. Except for the elementary education group, there was a predominance of male students in each group.

**Predictor Variables.** The total high school grade-point average (HSGPA) and two tests were used. The tests were part of the entrance battery of this university. The 1949 editions of the *American Council on Education Psychological Examination* (ACE) and the *Cooperative General Culture Test* (CGCT) were used. Subtest scores and total scores were tabulated. The Wherry-Doolittle method of test battery selection was used.

## Results

The most efficient single predictor of curriculum success was the HSGPA. In combination with the ACE Total score, it supplied the most efficient batteries, with an additional factor in the social science curriculum and two in the physical sciences.

The multiple correlations for the most efficient battery and each curriculum are shown in Table 1. Also shown are the respective Index of Forecasting Efficiency ( $E$ ), the Coefficient of Determination ( $R^2$ ), and the Standard Error of  $R$ .

The most efficient battery for predicting academic success in the commerce curriculum included the HSGPA and the ACE Total score. This battery accounted for 40.1 per cent of the criterion variance, compared to 35 per cent for the best single predictor (HSGPA).

These two factors, HSGPA and ACE Total, also comprised the most efficient battery for predicting success in the elementary education curriculum. This battery accounted for 53.4 per cent of the criterion variance, compared to 45 per cent for HSGPA, alone.

<sup>1</sup> This paper is a portion of a dissertation presented as partial fulfillment of the requirements of the degree of doctor of philosophy at the University of Utah. The writer is particularly indebted to Dr. F. B. Jex and Dr. R. D. Willey, who served variously as chairman of the dissertation committee.

Table 1

Multiple Correlations of Certain Predictor Variables and Success in Four Curricula at Brigham Young University

Curriculum	N	Battery	R	SE <sub>R</sub>	E	R <sup>2</sup>
Commerce	102	HSGPA & A.C.E. Total	.663	.060	23.4	40.1
Elementary Education	123	HSGPA & A.C.E. Total	.731	.042	31.8	53.4
Physical Science	133	HSGPA, A.C.E. Total	.733	.040	32.0	53.7
		CGCT Literature & General Science				
Social Science	78	HSGPA, A.C.E. Total	.507	.084	13.8	25.7
		CGCT General Science				

The factors, HSGPA and ACE Total, were supplemented by the CGCT Literature and General Science sub-tests in providing the most efficient battery for predicting success in the physical sciences. This battery accounted for 53.7 per cent of the criterion variance, compared to 33 per cent for HSGPA, alone.

The factor, CGCT Literature, dropped out of the above battery in providing the most efficient battery for predicting success in the social sciences, leaving the HSGPA, ACE Total, and CGCT General Science. This battery accounted for 26 per cent of the criterion variance, compared to 18 per cent for the ACE Linguistic score. It should be noted that the criterion reliability for this curriculum was substantially lower than that of the other curricula.

Multiple regression equations and conversion tables were prepared for each of the above batteries. It is possible for a counselor at Brigham Young University to take the student's HSGPA, ACE Total score, and CGCT Literature and General Science scores, and determine the predicted grade-point average (PGPA) for that student in any one or all of the four curricula studied.

#### Summary

1. The utilization of entrance test data and high school grade-point average provides the counselor at Brigham Young Uni-

versity with the basis for making differential predictions of academic success in four curricula.

2. For commerce and elementary education, the most efficient battery included the HSGPA and ACE Total scores. The respective R's were .633 and .731.

3. The physical sciences criterion was best predicted by a battery including the HSGPA, ACE Total, and CGCT Literature and General Science. R for this battery was .733.

4. The social science predictor battery included the HSGPA, ACE Total and CGCT General Science. R was .507.

5. The best single predictor was the HSGPA.

6. The reliability coefficients of the criterion measure (CGPA) clustered around .80 except for the social science curriculum with an r of .68.

Received June 2, 1953.

#### References

1. Crawford, A. B. and Burnham, P. S. *Forecasting college achievement*. New Haven: Yale University Press, 1947.
2. Monroe, W. S. (editor). *Encyclopedia of educational research*. New York: The Macmillan Company, 1950. Pp. 882-886.
3. Wallace, W. C. Differential predictive value of the A.C.E. Psychological Examination. *Sch. & Soc.*, 1949, 70, 23-25.
4. Wolf, R. R., Jr. Differential forecasts of achievement and their use in educational counseling. *Psychol. Monogr.*, 1939, 51, 1-53.



# Performance of College Students on a Mechanical Knowledge Test

Benjamin Balinsky and Charles Hujisa

*City College of New York*

When given the SRA Mechanical Aptitude Test as part of a course in Vocational Psychology, students commented that they did not do as well on the Mechanical Knowledge subtest as on the Space Relations and Shop Arithmetic. In order to test the comment, the SRA Mechanical Aptitude Test results of 112 male students were tabulated. All students were in either the junior or senior year of the School of Business of the City Col-

lege of New York and between the ages of 19 and 22.

The Revised Minnesota Paper Form Board Test was available on the 112 students, and the Location, Blocks, and Pursuit subtest scores of the MacQuarrie Test for Mechanical Ability on 50 of the 112 students. These tests were also included in the study.

Test intercorrelations were calculated for all combinations of tests. Tests of signifi-

Table 1  
Test Intercorrelations for College Students †

	Mech. Knowl.	Space Rel.	Shop Arith.	Total SRA	Rev. Minn. P.F.B.	Loc.	Bl.	Pur.
Mech. Knowl.	—	.09	.07	.64**	.19*	.20	.28*	.23
Space Rel.		—	.18*	.50**	.36**	.18	.42**	.44**
Shop Arith.			—	.44**	.11	.30*	.14	.33*
Total SRA				—	.36**	.31*	.39**	.44**
Rev. Minn. P.F.B.					—	.17	.52**	.50**
Location						—	.13	.24*
Blocks							—	.32**
Pursuit								—

† (Total SRA Minus Mech. Knowl. × Mech. Knowl.) = .12

† (Total SRA Minus Space Rel. × Space Rel.) = .18\*

† (Total SRA Minus Shop Arith. × Shop Arith.) = .15

† The correlations with the Location, Blocks and Pursuit subtests are based on 50 subjects; others on 112.

\* Significant at the 5% level.

\*\* Significant at the 1% level.

Table 2

Means, Standard Deviations and Tests of Significance for College and SRA Male Trainee Groups

	College Group		SRA Male Trainees		t	P
	Mean	S.D.	Mean	S.D.		
Mech. Knowl.	25.0	6.3	31.8	7.1	10.27	<.001
Space Relations	20.4	4.1	19.0	4.7	3.19	<.01
Shop Arith.	14.4	3.7	9.8	3.8	11.90	<.001
Total	59.6	10.6	60.5	12.2	0.78	.42

cance were computed for the difference between the means of the test scores of the college students and the norm group of male trainees in the SRA Mechanical Aptitude Test. These data are presented in Tables 1 and 2.

Incidentally, the mean Mechanical Knowledge score of the students is at the 19th percentile of the male trainee norms, the mean Space Relations at the 55th percentile, and

the mean Shop Arithmetic at the 87th percentile. The mean score of the college students on the Revised Minnesota Paper Form Board is at the 70th percentile of the machine and electrical apprentice applicants norms and the difference between the means of both groups is significant at  $< .001$  in favor of the students.

*Received June 19, 1953.*

## Relation of Scholastic Aptitude to Socioeconomic Status and to a Rural-to-Urban Continuum

Norman F. Washburne and Dean C. Andrew

*Southern State College, Magnolia, Arkansas*

Scholastic aptitude tests play an important role in modern education. They are used in schools for the purpose of aiding students in selecting courses and vocations, and have a variety of other uses in guidance, counseling, and various aspects of research. At Southern State College it is the practice to administer the college level ACE *Psychological Examination* to all entering freshmen. Where such a practice prevails, it is desirable to know if the test measures what it is intended to measure, and what factors, if any, might bias the results of the test.

The student body of Southern State College is composed primarily of residents of southern Arkansas, northern Louisiana, and eastern Texas. It is a regionally homogeneous population. There are no significant immigrant groups, and the undergraduate students are all Caucasians. However, the population is quite varied in two respects: (1) the socioeconomic status of the individual student; and (2) the sizes of the communities in which the students have grown up. The question therefore arises, do these variations in socioeconomic status and degree of urbanization affect the scholastic aptitude of the students as measured by the ACE *Psychological Examination*?

In order to test this question, a sample of 100 students was drawn at random from those who had been enrolled in April 1952.<sup>1</sup> These students had been given the ACE and had also filled out a questionnaire which made it possible to determine their socioeconomic statuses and the relative degrees of urbanization of their residence histories. Coefficients of total and partial correlation were computed in order to determine the relationships among the three variables.

*The Scholastic Aptitude Test.* The ACE

<sup>1</sup> This sample size approximates one-sixth of the undergraduate student body at the time.

*Psychological Examination*<sup>2</sup> is an instrument designed to measure the scholastic aptitude of American college freshmen. Its scoring yields three measures: the Q score; the L score; and the total score. The Q score is a measure of the respondent's ability to solve problems of quantitative nature. The range of the Q scores of our sample was from 9 to 60. The L score is the measure of the respondent's ability to solve problems of a linguistic nature. The range of the L scores of our sample was from 27 to 92. The total score is a sum of the Q and L scores and is a measure of the total scholastic aptitude of the respondent. The range of the total scores of our sample was from 36 to 142.

*The Socioeconomic Status Scale.* The socioeconomic status scale is an instrument designed to quantify the social and economic position of the college student.<sup>3</sup> Its scores are based upon the occupation of the student's father, and upon the educational attainments of both of the student's parents. The occupational and educational factors are weighted equally. The scale differs from some other socioeconomic status scales used in similar studies in two ways: (1) its occupational factor is not arbitrarily scored, but rather is based upon the students' own evaluation of occupations representative of those of their fathers; and (2) it does not assume that social classes exist as discrete

<sup>2</sup> American Council on Education Psychological Examination, 1948 College Edition, Educational Testing Service, Cooperative Test Division (Princeton, New Jersey), 1951.

<sup>3</sup> Details of the construction and validation of the Socioeconomic Status Scale and the Residence History Scale are to be found in Norman F. Washburne, "Urban" Attitudes and Responses as Related to Residence in Urban Communities and to Socioeconomic Status, Ph.D. Dissertation, Washington University, St. Louis, Missouri, 1953. This work has also been published in mimeographed form as an Institutional Study of Southern State College, Magnolia, Arkansas, and a limited number of copies are available on request.



Table 1

Total and Partial Correlation of Scholastic Aptitude Scores with Socioeconomic Status Scores,  
100 Southern State College Students

Coefficient of Total Correlation		Probability of Null Hypothesis	Partial Correlation with Residence History Held Constant	Probability of Null Hypothesis
Q score	.024	P > .05	.024	P > .05
L score	.123	P > .05	.070	P > .05
Total score	.166	P > .05	.115	P > .05

$r$  of Socioeconomic Status vs. Residence History =  $+.19$ .

cultural units, but rather assumes a continuum of socioeconomic statuses. The points of the scale are handled statistically as if they were the midpoints of intervals along the continuum. The theoretical as well as the actual range of the socioeconomic status scale is from 2 to 10.

To clarify the meaning of the scale the following examples are offered: The father of one student scoring 10 on the socioeconomic status scale is an owner of a large manufacturing plant. Both of the student's parents had gone to college and one of them had taken graduate work beyond the baccalaureate degree. On the other end of the scale the father of one student who scored 2 is a day laborer on a farm. Neither of this student's parents had completed the sixth grade in school.

The occupational factor of the scale correlated highly with North and Hatt's similar scheme<sup>4</sup> and so is judged to be valid. Socioeconomic status scores were computed for a sample of 100 students from data gathered on two different occasions, and the scale was found to be reliable.

**Residence History Scale.** The residence history scale is an instrument designed to quantify the degree of urbanization of the backgrounds of individuals. It is, as far as we know, the first instrument which goes beyond the simple characterization of students' home-towns as being either rural or urban. It is a complicated device which takes into account the size, degree of isolation, and

proximity of larger urban centers of all the individual's places of residence from the time he entered the first grade until the present. It also takes into account the length of time the individual spent in each place of residence. The residence history scale assumes a rural-to-urban continuum. It has a theoretical range from 0 to 50. A score of 0 would indicate that the individual has lived all his life more than 100 miles away from the nearest community of 250 population. At the other extreme, a score of 50 would indicate that the student has lived all his life within 6 miles of a city of a population of at least one-half million. The actual range of the residence history scores of our sample was from 10 to 48. Residence history scores were computed for 100 students from data gathered on two different occasions, and the scale was found to be reliable.

### Results

The relationships between the scores on the scholastic aptitude test and the socioeconomic status scores of the sample are presented in Table 1.

It can be seen from Table 1 that all coefficients of total correlation were low and statistically not significant. However, since the coefficient of total correlation between residence history scores and socioeconomic status scores was found to be  $+.19$ , it seemed feasible to seek an understanding of the effects of each of the factors upon the scholastic aptitude scores when the other was held constant. Table 1 therefore also presents the coefficients of partial correlation of the test

<sup>4</sup> Cecil C. North and Paul K. Hatt. Jobs and occupations: a popular evaluation. *Opinion News*, September 1, 1947, pp. 3-13.

Table 2

Total and Partial Correlation of Scholastic Aptitude Scores with Residence History Scores,  
100 Southern State College Students

	Coefficient of Total Correlation	Probability of Null Hypothesis	Partial Correlation with Socioeconomic Status Held Constant	Probability of Null Hypothesis
Q score	.245	P < .01	.245	P < .01
L score	.302	P < .01	.286	P < .01
Total score	.308	P < .01	.295	P < .01

scores with socioeconomic status scores while the residence history scores are held constant. None of the resulting relationships are shown to be significant.

The relationships between residence history scores and the scholastic aptitude scores are presented in Table 2.

All coefficients of total correlation between residence history scores and the scholastic aptitude scores are shown in Table 2 to be significant at the one per cent level. That means that the relationship between the factors would happen less than one time in a hundred by chance. When coefficients of partial correlation of the scholastic aptitude scores with the residence history scores were calculated while socioeconomic status scores were held constant, all of the resulting coefficients were slightly lower than the total coefficients with the exception of the Q score which remained the same. However, even these slightly lower coefficients of partial correlation were found to be significant at the one per cent level. All of the relationships were in the direction of rural-to-urban, i.e., the more urban the background the greater the scholastic aptitude.

### Summary and Conclusions

This investigation attempted to discover the relationships between scholastic aptitude, socioeconomic status, and placement of the individual upon a rural-to-urban continuum, as these variables applied to Southern State College students. The results seem to justify the following conclusions:

1. For this group of college students there is no significant relationship between socioeconomic status and scholastic aptitude as measured by the ACE *Psychological Examination*.

2. There is a significant, though low correlation between placement of the students' residence history upon a rural-to-urban continuum, and their scholastic aptitude as measured by the ACE *Psychological Examination*. That is to say that students from more urban backgrounds tend to receive higher scores than do students from rural backgrounds.

Because these findings apply only to Southern State College students, it is suggested that further research be conducted on students in other schools and in other regions, to see if the findings are confirmed.

Received May 14, 1953.

## Further Results on Group Manual Dexterity in Men

Andrew L. Comrey and Gerald Deskin

*The University of California at Los Angeles*

In a previous experiment,<sup>1</sup> 65 pairs of volunteer male university students were given six individual trials on the Purdue Pegboard, Assembly Task, and six trials on the Assembly Task with the two members of each pair working together on the same assemblies rather than individually on separate boards. The members of each pair were divided on the basis of the total of the last four individual trials, Assembly Task, into "high" and "low" categories. Reliabilities were determined for "high," "low," and "group" performances, using alternate trials and correcting for doubled length. Correlations of the "high" and "low" performances with the "group" performance and with each other were computed and corrected for attenuation. The multiple correlation and regression weights were obtained for predicting "group" performance from "high" and "low" individual performances. The results showed that less than half the group performance variance could be predicted from a knowledge of the individual performances, even with the effects of errors removed. The level of group performance was only slightly more dependent on the "low" individual performances. For all practical purposes, equal weights could have been used for "high" and "low" scores in predicting "group" performance.

The present experiment was designed to provide a check on the first experiment and to determine the effect of an alteration in the nature of the individual task on the amount of group variance which could be predicted. One of the hypotheses offered to account for the fact that much of the variance in the group performance scores could not be predicted from a knowledge of the individual performances was that the two tasks might have been too unlike each other. Although

the same end product resulted in both individual and group performance, the latter required the subjects to alternate the operations they performed on successive assemblies. The first subject, for example, would place a peg in the first hole on his side of the board, after which the second subject would add a washer and the first subject would follow with a collar, and finally the second subject would complete the assembly with another washer. Instead of repeating this operation, however, the subject who finished the assembly would begin the next assembly by placing a peg in the second hole on his side. The first subject would then place on the first washer, and so on.

Since in the individual task, the subject performed each assembly just like the previous one, he was not confronted with the additional task of altering his set for each subsequent assembly, as he was required to do for the group performance task. It was hypothesized that this requirement for changing set might have introduced additional abilities into the task which were not present in the individual task, thereby lowering the validity of the individual performance scores for predicting the group performance scores.

To test this hypothesis, the experiment was repeated using a redesigned individual task which required the subjects to make a change of set on each assembly like that to be required later in the group task. Instead of using the standard instructions for the Purdue Pegboard, Assembly Task, the subjects were instructed to begin each assembly after the first one with the same hand used to place the final washer on the preceding assembly. In this way, the subject was required to make alternate assemblies with reversed hand operations, substituting the left hand for those operations previously performed with the right hand, and vice versa.

<sup>1</sup> Comrey, A. L. Group performance in a manual dexterity task. *J. appl. Psychol.*, 1953, 37, 207-210.



Table 1  
Summary of Results

Score	M	s	$r_{11}$	Corrected $r$ with			Beta Weight
				High	Low	Group	
High	156 (192)	18.0 (16.5)	.89 (.90)	1.00	.50 (.52)	.55 (.56)	.31 (.35)
Low	137 (173)	17.2 (16.8)	.94 (.92)	.50 (.52)	1.00	.64 (.59)	.49 (.41)
Group	186 (178)	19.2 (19.2)	.77 (.87)	.55 (.56)	.64 (.59)	1.00	
	$R = .69$ (.66)			$R^2 = .48$ (.44)			

### Results and Discussion

In every way except those differences already mentioned, the experimental procedure and treatment of the data were the same as for the first experiment,<sup>2</sup> and therefore will not be repeated here. The sample was made up entirely of undergraduate men this time, whereas about one-third of them were graduate students before; the previous 65 pairs of subjects was reduced to 47 pairs for this experiment. The results are summarized in Table 1. The figures included in parentheses show the results from the first experiment while the numbers immediately above them are the corresponding values in the present research.

In the first column of Table 1 are listed the total score categories, "high," "low," and "group," standing, respectively, for those total performances already described. The means and standard deviations of the three sets of scores are given in the second and third columns, respectively. These are based on the totals of the last four of six trials. This procedure was used to obtain greater stability. In the fourth column are given the split-half reliability estimates for the three types of scores. The next three columns of Table 1 give the intercorrelations of the total score variables, corrected for attenuation in both variables. The last column contains the beta weights for predicting "group" performance from "high" and "low" individual

performances. The multiple correlation,  $R$ , and  $R^2$ , are given in the bottom row of the table. Both the beta weights and  $R$  were computed using the corrected correlations. The uncorrected correlations for the present experiment were: low-high, .46, low-group, .53, and high-group, .47. For the previous experiment, the corresponding uncorrected correlations were .48, .53, and .50.

An inspection of Table 1 reveals certain discrepancies which require some comment. In the present experiment, the mean individual scores were considerably lower than for the previous experiment, which was expected because the task was more difficult. This resulted in a slight increase in variance, too, which again could have been expected. The present group performance mean was only slightly higher, probably due to individual-task practice in changing set, not available to performers in the first experiment. The variances were identical for the group task in both experiments, an outcome consistent with expectations in that the task was exactly the same in both cases.

The reliabilities compare very favorably in the two experiments, except for the group performance score. In this case, the present figure was lower than the previous one. The corrected correlations are close in the two experiments, although the discrepancies are in opposite directions for the low-group and high-group correlations, resulting in a more impressive difference between the beta weights.

<sup>2</sup> Comrey, A. L., *op. cit.*

Whereas the beta weights were fairly close in the first experiment, the low scores emerged in this research with a definite edge for predicting group performance, although the difference was still short of statistical significance.

Looking at the comparative multiple correlation coefficients and their squares, it is evident that no startling improvement has occurred in the amount of group-performance-score variance which can be predicted from a knowledge of individual performance scores. The proportion of predicted variance is still less than half. This figure was achieved only through using correlations corrected for attenuation. The proportion of variance practically predictable would be less. It is perhaps worth mentioning that the multiple *R* values would have been even closer if the "group" score reliability in the second experiment had been higher. Since the figure obtained may be spuriously low, it would be well to consider the gain actually achieved with some caution.

The results do not bear out the hypothesis entertained that prediction of group performance scores can be increased markedly by making the individual task apparently more like the group task in the actual operations.

Two other hypotheses, as yet untested, were offered in the previous article to account for the additional unpredicted variance. The group task may involve some special traits introduced by the necessity of cooperating with another person and there may be interaction effects among individuals over and above stable trait influences. Attempts will be made in further work to explore the nature of this as yet unpredicted variance.

#### Summary

A previously reported experiment was repeated with an altered design to test the former results and a hypothesis offered to account for the fact that group performance scores on a manual dexterity task could only be predicted rather imperfectly from knowledge of individual scores on a similar task. The hypothesis was offered that the prediction might be substantially improved by a change in design to make the group and individual tasks more comparable in the character of the operations involved. The amount of improvement in prediction obtained was so slight as to require the rejection of the hypothesis.

*Received May 8, 1953.*

## Effects of Fatigue and Anxiety on Certain Psychomotor and Visual Functions<sup>1</sup>

Sherman Ross, T. A. Hussman, and T. G. Andrews

University of Maryland

This experiment was an attempt to investigate the degree of behavior decrement produced by the experience of fatigue and threat of bodily damage occasioned in the competitive athletic sport of boxing. The dependent variables chosen as possible indicators of behavior decrement were: (a) steadiness score; (b) body sway score; (c) body sway time score; (d) tapping rate; and (e) critical flicker frequency. The primary purpose of the experiment was to determine whether or not performance on each of the five dependent variables changes significantly as a result of intensive muscular exercise (fatigue) or the fear of bodily injury (anxiety) or the interaction of these conditions in the collegiate competitive boxing situation.

There has been some speculation in the past regarding the damaging effects on behavior of sustained head blows such as received in continuous training in boxing (13). In addition to these interests in boxing, such a situation appears to offer a realistic condition of systemic fatigue, high motivation, and anxiety such as could not be attained under the usual conditions of laboratory investigations. These characteristics are not unlike those which obtain in certain field conditions of military operations and combat. In the general search for indicators of behavior decrement for military purposes, the use was made of boxing behavior to approximate these characteristics of military importance.

The basis for the selection of the indicators used in this investigation is described below for each of the five dependent variables together with a description of the manner of testing.

<sup>1</sup> This experiment is one of a series of studies on behavior decrement performed under Contract No. DA-49-007-MD-222 between the Medical Research and Development Board, Office of The Surgeon General, Department of the Army and the University of Maryland. The opinions and assertions expressed in this report do not necessarily reflect the views of the Department of the Army.

### Tests and Indicators Used

*Steadiness* has been demonstrated to show changes under certain conditions of stress, and it has been reported to change with fatigue or work output (1, 4, 5, 18). Hand steadiness and tremor have also been related to emotional stimulation (6, 7) and to certain conditions of motivation (4). Because of these features, a test of hand steadiness was included among the dependent variables. For this test a target hole in a vertically adjustable metal plate was used. The subject's task was to keep a 0.02 inch diameter stylus inserted into the 0.136 inch hole for 20 seconds with the arm fully extended and unsupported. The number of contacts with the edge of the hole during this period served as the score.

*Body sway* measurements have offered rather controversial results in the past when related to fatigue (11, 18) and to loss of sleep (5, 15). Because of the possible effects of head blows sustained in boxing, measures of body sway were obtained. For this purpose an arrangement similar to that for steadiness was used. However, in this case the stylus was longer and the hole diameter was 0.358 inch. The subject was required to hold the stylus in the hole, but in this case without the aid of visual cues. When contact was made with the edge of the hole, a buzzer was automatically sounded as a signal to the subject. Two scores were derived from this test: a *body sway score* of the number of contacts made in the 20 second period, and a *body sway time score* consisting of the total amount of time in seconds the stylus was in contact with the edge of the hole during the observation period. These were treated as separate scores in the analysis of the data.

*Tapping* tests serve as measures of rather simple performance, but have been considered by some investigators as useful indices of fatigue (15, 16). Tapping has been shown to be related to the decrement produced by high altitude (9). The tapping test apparatus used here consisted of the Dunlap modification of the Whipple Tapping Board (3) and a 0.20 inch diameter stylus. The tapping targets were two 3 inch square brass plates separated by 1 inch of bakelite. The subject was to tap alternately on the plates as rapidly as possible for a period of 15 seconds. The score used was the total number of taps on the plates in this allotted time. This brief time period was used as an attempt to diminish the factor of learning, which has been shown to affect tapping scores (17).



*Critical Flicker Frequency* has been used in several investigations on fatigue with controversial results (12, 14, 18). There has been some indication that CFF changes when the individual is subjected to intensive strain (2). The apparatus used in the present study was the Krasno-Ivy Flicker Photometer (8), which is essentially an episotister arrangement delivering square wave flashes of light on a  $\frac{3}{4}$  inch ground glass screen. The subject was seated 5 feet from the stimulus screen. A modified method of limits was used, in which the experimenter manipulated the stimulus from "fusion to flicker" and the subject responded at his threshold. Six "descending" trials were employed, the first two serving as practice. The score or threshold measure was the mean number of flashes for the last four trials.

In each of the above tests only a brief period could be devoted to obtaining a score, since in many instances the subjects were being measured immediately after strenuous exercise and before they were covered, rubbed down, or bathed. Longer testing periods would have increased the reliabilities of the measures taken, but also might possibly have allowed the injurious effects of chilling the subjects.

### Subjects

Twenty-four male college students ranging in age from 19 to 25 years were used as subjects. Twelve of the group were experienced collegiate boxers and members of the University of Maryland Boxing Team for 1952. The remaining subjects were members of a Physical Education class in boxing and should be classed as novice boxers. All subjects were in excellent physical condition.

### Independent Variables and Experimental Design

Each of the 24 subjects was measured three times on each of the tests under each of four conditions of the investigation. These four conditions were as follows:

- At rest, no previous strenuous exercise, no expectation of going into the ring to fight.
- Before fighting a three-round supervised bout, no previous exercise.
- After three rounds of very strenuous work-out on a heavy punching bag, not in the ring nor expecting to go into the ring.
- After fighting a three-round supervised bout with an opponent.

These four conditions yield a basic  $2 \times 2$  block of the experimental design, which is diagrammed in Table 1. It may be seen that this arrangement opposes the no-exercise conditions (F-O) to the heavy exercise conditions (F) for a test of the change in each variable as a result of fatigue. The test of change in each variable due to the anxiety occurring in the boxing situation

is made by opposing the no-anxiety conditions (A-O) to the high anxiety conditions (A). The problem of fatigue in this arrangement is quite straightforward. The problem of anxiety, however, offers some question. In this regard it may be said that all observations on and reports from the men immediately before and after such competitive boxing indicate severe tension and concern over the threat of pain and bodily damage or loss of the bout.

In order to minimize the effects of the order of taking the tests in the battery, each subject was randomly assigned to one of the 24 possible orders of test administration, which he maintained throughout the experiment. Each subject was measured 12 times on each test, three times under each of the four experimental conditions. The restriction placed upon the order of the conditions was that the first time a subject took the tests he was under the rest condition so that giving the instructions did not interfere with the condition nor the reverse.

### Results and Discussion

The results are presented and analyzed separately for each of the dependent variables studied. In each case reference is made to the paradigm presented in Table 1, and the code letters used refer to the designated experimental conditions and their combinations as a system for presenting the obtained means.

The experiment was conceived and designed to allow analysis of the results in two separate manners. The fact that each block of measures taken on the twenty-four Ss is replicated twice allows the use of a within-individual estimate of variance to be used as an error

Table 1  
Experimental Design Indicating the Conditions of Measurement and Their Relationships

		FATIGUE		
		Absent	Present	
ANXIETY	Absent	n = 24 j = 3	n = 24 j = 3	A-O
	Present	n = 24 j = 3	n = 24 j = 3	A
		F-O	F	

term to evaluate the effects of the treatment conditions on the variables in the population used. This error term contains variance of two types, that associated with instrument error and individual diurnal variation. This analysis is intended to test the theoretical and perhaps somewhat obvious question of whether these variables are affected by the treatment conditions of fatigue and anxiety in the sample used.

The second analysis, which uses an estimate of the individual differences variance as the error term, is intended to answer the question of whether these test variables are useful as reliable indices of the independent variables for practical application. Frequently the question of whether a variable changes significantly as a result of such conditions as fatigue and anxiety has been confused with the question of whether it may be used as an adequate indicator of these conditions. The two analyses employed test each of these questions in turn with what is felt to be the proper error term for each. The second analysis also provided a test of the replications as a main effect, thus enabling a check on possible changes due to learning, the presence of which of course would cast some question on their usefulness as indicators. In all cases tests of homogeneity of variance were satisfied. Table 2 presents the means for each experimental condition for each of the dependent variables used, according to the paradigm in Table 1. Tables 3 and 4 present composite results of the tests of significance. Reference is made to these three tables in the description of results for each type of experimental measure.

**Steadiness.** The total mean score for all subjects under all conditions was 72.23; for condition F-O = 62.0, F = 82.46, A-O = 72.25, A = 72.22. The differences associated with fatigue conditions are significant at the .001 level, as are individual differences. Anxiety conditions effected no change in the measures.

There is a questionable interaction between fatigue and anxiety, and the interaction between fatigue and individual differences is very significant as is the interaction of anxiety and individual differences. From these combinations of interactions it appears that anx-

Table 2

Means of Experimental Results for Specified Tests and Conditions

		Fatigue		
		0	+	
Steadiness	Anx.	0	60.44    84.05	72.25
		+	63.57    80.86	72.22
			62.00    82.46	72.23
		0	+	
Body Sway	Anx.	0	26.25    32.11	29.18
		+	28.22    34.48	31.35
			27.24    33.30	30.26
		0	+	
Body Sway Time	Anx.	0	590.36    638.24	614.30
		+	490.89    667.70	579.30
			540.62    652.97	596.80
		0	+	
Tapping	Anx.	0	77.54    83.96	80.75
		+	81.75    82.62	82.18
			79.64    83.29	81.47
		0	+	
CFF	Anx.	0	48.450    49.471	49.005
		+	48.570    47.811	48.190
			48.555    48.641	48.598

ity may act here to increase the scores of some individuals and decrease or not affect the scores of others, thus destroying the main effect. Anxiety then may be acting as a sensitizer to fatigue effects in some instances and a desensitizer in other instances. No significant change was observed in successive

Table 3

Analyses of Variance for the Specified Experimental Variables, Using "Within Individuals" as Measure of Experimental Error<sup>1</sup>

Source	df	MS for Steadiness	MS for Body Sway	MS for Body Sway-Time	MS for Tapping	MS for CFF
Fatigue Conditions	1	30,114.67***	2,628.12***	908,664.34***	957.03***	.51
Anxiety Conditions	1	.09	333.68**	88,235.00	148.78*	47.61***
Individuals	23	3,048.33***	341.42***	435,366.00**	730.70***	149.85***
Interactions:						
Fat. $\times$ Anx.	1	718.83*	3.56	299,215.59*	552.78***	51.77***
Fat. $\times$ Ind.	23	322.07***	64.24*	84,111.68	29.16	5.48***
Anx. $\times$ Ind.	23	301.88**	68.32**	54,416.36	94.64***	4.68**
Fat. $\times$ Anx. $\times$ Ind.	23	191.63	77.48**	141,400.42***	59.87	3.30
Error:						
Within Individuals (replications)	192	143.14	34.66	56,909.71	38.17	2.41
	287					

<sup>1</sup> The asterisks identify the conventional levels of significance: \* for .05, \*\* for .01, and \*\*\* for .001.

measurements on the same individual under the same condition. The general conclusion here is that fatigue produces a general decrease in steadiness.

**Body Sway.** The total mean score for all conditions was 30.26; for condition F-O = 27.24, F = 33.3, A-O = 29.18, A = 31.35. Fatigue effects very significantly increase

body sway, and anxiety appears also significantly to produce the same results. Individual differences are also significant here. The interactions were insignificant when compared with the highest order interaction, as recommended by McNemar (10). No significant effects were obtained for repeated measures under the same conditions. The

Table 4

Analyses of Variance for the Specified Experimental Variables, Using "Within Cells" as Measure of Experimental Error<sup>1</sup>

Source	df	MS for Steadiness	MS for Body Sway	MS for Body Sway-Time	MS for Tapping	MS for CFF
Fatigue Conditions	1	30,114.67***	2,628.12***	908,664.34***	957.03**	.51
Anxiety Conditions	1	.09	333.68*	88,235.00***	148.78	47.61
Replications	2	338.04	21.26	128,191.06***	354.59*	.76
Interactions:						
Fat. $\times$ Anx.	1	718.83	3.56	299,215.59***	552.78*	51.77
Fat. $\times$ Repl.	2	3.96	32.04	45,262.12**	3.94	7.84
Anx. $\times$ Repl.	2	216.58	60.59	20,433.86	8.53	2.41
Fat. $\times$ Anx. $\times$ Repl.	2	34.17	5.59	12,374,698.76***	61.86	5.52
Error:						
Within Cells (individual differences)	276	417.28	69.20	8,120.64	99.64	15.16
	287					

<sup>1</sup> The asterisks identify the conventional levels of significance: \* for .05, \*\* for .01, and \*\*\* for .001.



general conclusion here is that fatigue and anxiety both increase body sway and significantly more for some individuals than for others. The results obtained serve to corroborate other studies on steadiness and body sway (1, 4, 5, 6, 7, 11, 18).

*Body Sway Time Scores.* The total mean score for all conditions was 596.80; for condition F-O = 540.62, F = 652.97, A-O = 614.30, A = 579.30. Fatigue effects are very reliable in their action to increase these scores. However, replication measures under the same conditions as well as differences among individuals were also highly significant. Because of these features and a very significant triple interaction effect, there was judged to be such a large amount of uncontrolled variability that no definite conclusions are offered for this measure of behavior decrement.

*Tapping.* The total mean score for all conditions was 81.47; for condition F-O = 79.64, F = 83.29, A-O = 80.75, A = 82.18. As in the case of the body sway time scores, there is a large amount of uncontrolled variability evidenced for the tapping measures. The fatigue condition acted to increase tapping reliably. However, replications within the same conditions as well as individual differences proved significant. There is probably too great a learning factor allowed in the conditions of measurement of tapping. The results in general indicate that the tapping test would be a sensitive indicator of behavior decrement if the learning factor were better controlled.

*Critical Flicker Frequency.* The total mean score for all conditions was 48.598; for condition F-O = 48.555, F = 48.641, A-O = 49.005, A = 48.190. Fatigue alone did not produce any significant change, but anxiety effects were highly significant in their decrease of CFF. Individual differences were significant, and replication effects were not significant.

Examination of the significance of the interactions in the case of CFF suggests that the same relationship holds between fatigue and CFF that obtained between anxiety and steadiness. The interaction between anxiety and individuals is significant, indicating a differential effect. The interaction of fatigue

and individuals is also present and suggests that some individuals change in one direction here while others change minimally or in the other direction, thus reducing the main effect that is predictable from fatigue. A more important interaction is found between the main effects of fatigue and anxiety. From the specific results obtained with CFF, it appears that this test may be a useful one for studies on behavior decrement only in situations of individual cases.

*Interrelationships Among the Measures.* Intercorrelations were obtained among the average scores of the tests as they appeared under the rest or control condition. These Pearson correlations were obtained only on the 24 Ss, and it was found that only two such correlations were significant. These were the correlations between steadiness and body sway ( $r = .550$ ) and between steadiness and the body sway time scores ( $r = .407$ ).

Body sway appears to have a factor in common with steadiness, and this is possibly the reason that body sway measures were found to be adequate indices of the stress involved in this investigation. It is also possible that the body sway test involves a factor or factors not present in the steadiness test, because the former was found to be a significant indicator of anxiety effects, while this did not hold for the steadiness test.

There was one source of variation that was impossible to control, namely the actual amount of bodily damage or physical punishment sustained by each of the Ss during the conditions of competitive boxing. It was felt that some system should be instituted that would allow a possible check on the validity of some of the experimental assumptions, and so correlations were computed between the number of head blows received and scores on each of the tests. The estimates of head blows were furnished by Mr. Frank Cronin, the University Boxing Coach, who observed every bout and tallied blows on a prearranged data form. None of the correlations was found to be statistically significant on a one-tail  $t$  test, which is the appropriate test considering the hypothesis in this case. It would appear that within the limits of the measuring techniques and the design of the

study, the number of head blows sustained had little or no effect on the test scores of the Ss.

As part of another investigation, to be reported elsewhere, protracted boxing experience with its attendant number of head blows produced no reliably indicated changes in the electroencephalographic records of amateur boxers, some of whom were from among the Ss used in the present investigation.

As far as the present results are concerned, it appears that measures of steadiness more than the other variables tested satisfy more of the criteria of reliability and predictability to be used as indicators of behavior decrement. Hand steadiness serves for indications of fatigue, and body sway which is a form of steadiness measure serves for indication of either fatigue or the type of anxiety produced in this study. These suggestive results may be taken as recommendations for further investigations under a greater variety of stress conditions.

The other variables employed in this study may be made into more useful measures for studies of stress if their trial-to-trial variation and very wide individual differences may be diminished by deriving scores through other techniques, reducing practice effects, and otherwise accounting for the larger relative amounts of variability now classifiable as experimental error.

### Summary and Conclusions

As part of a larger research program on indicators of behavior decrement, this experiment investigated the comparative value of several selected measures of behavior decrement under conditions of fatigue and anxiety. The dependent variables chosen as possible indicators of behavior decrement were: (a) steadiness; (b) body sway; (c) body sway time score; (d) tapping rate; and (e) critical flicker frequency. The primary purpose of the experiment was to determine whether or not performance on each of the five dependent variables changed significantly as a result of intensive muscular exercise (fatigue) or the fear of bodily injury (anxiety) or the interaction of these conditions in the collegiate competitive boxing situation.

Twenty-four boxers were measured under the following four conditions: at rest; after heavy exercise; before fighting; and after fighting. The tests were administered three times to each subject under each of the experimental conditions. The analysis of variance technique was used to test the changes in each variable as a function of the independent variables. Two separate analyses of the results were made: (1) using "within individuals"; and (2) using "within cells" as the measure of experimental error. The results permit the following major conclusions:

1. Hand steadiness scores decreased significantly with fatigue, but not with the anxiety conditions. No significant change was observed in successive testing on the same individual under the same conditions.

2. Fatigue and anxiety significantly increased body sway scores.

3. Body sway time scores were found to be unreliable, although the fatigue conditions significantly increased these scores.

4. Tapping was found to be unreliable, possibly due to a learning factor. Significant changes were found, however, in the test scores as a result of both fatigue and anxiety.

5. Critical flicker frequency thresholds were shown to decrease significantly as a result of anxiety. The reliability of the test was high, and it is felt that it may be useful in studies of behavior decrement in situations of individual cases.

6. No relationship was found between the dependent variables used and the number of head blows received by the subjects during a boxing bout.

7. Measures of steadiness more than the other variables tested satisfy the criteria for indicators of behavior decrement. Hand steadiness serves as an indicator of fatigue, and body sway (which is a form of steadiness measure) may serve as an indicator of either fatigue or the type of anxiety produced in the experiment. The remaining variables tested in this experiment may be made into more useful measures for studies of the effects of stress if trial-to-trial variation and the very wide individual differences exhibited are diminished.

*Received June 8, 1953.*

## References

1. Bousfield, W. W. The influence of fatigue upon tremor. *J. exp. Psychol.*, 1932, 15, 104-107.
2. Brozek, J. and Keys, A. Flicker fusion frequency as a test of fatigue. *J. indust. Hyg.*, 1944, 26, 169-174.
3. Dunlap, K. Improved form of steadiness tests and tapping plate. *J. exp. Psychol.*, 1921, 4, 430-433.
4. Eaton, M. T. The effect of praise, reproof and exercise upon muscular steadiness. *J. exp. Educ.*, 1933, 2, 44-59.
5. Edwards, A. S. Effects of the loss of one hundred hours of sleep. *Amer. J. Psychol.*, 1941, 54, 80-91.
6. Edwards, A. S. Finger tremor and battle sounds. *J. abnorm. soc. Psychol.*, 1948, 43, 396-399.
7. Kellogg, W. N. The effect of emotional excitement upon muscular steadiness. *J. exp. Psychol.*, 1932, 15, 142-165.
8. Krasno, L. R. and Ivy, A. C. The response of the flicker fusion threshold to nitroglycerin and its potential value in the diagnosis, prognosis, and therapy of subclinical and clinical cardio-vascular disease. *Circulation*, 1950, 1, 6, 1267-1276.
9. Malmø, R. B. and Finan, J. L. A comparative study of eight tests in the decompression chamber. *Amer. J. Psychol.*, 1944, 57, 389.
10. McNemar, Q. *Psychological statistics*. New York: John Wiley and Sons, 1949. P. 288.
11. Ryan, A. H. and Warner, M. The effects of automobile driving on the reactions of the driver. *Amer. J. Psychol.*, 1936, 48, 403-421.
12. Simonsen, E. and Enzer, E. Measurement of fusion frequency of flicker as a test of fatigue of the central nervous system: observations on laboratory technicians and office workers. *J. indust. Hyg.*, 1941, 23, 83-89.
13. Steinhaus, A. H. Boxers brains swapped for medals. *J. of the Amer. Assn. for Health, Physical Ed. and Recreation*, 1951, 8, 12-14.
14. Tyler, D. B. The fatigue of prolonged wakefulness. *Fed. Proc.*, 1947, 6, 218.
15. Warren N. and Clark B. Blocking in mental and motor tasks during a 65-hour vigil. *J. exp. Psychol.*, 1937, 21, 97-105.
16. Wells, F. L. A neglected measure of fatigue. *Amer. J. Psychol.*, 1908, 19, 345-358.
17. Wells, F. L. Normal performances on the tapping test before and during practice, with special reference to fatigue phenomena. *Amer. J. Psychol.*, 1908, 19, 437-483.
18. Wulfeck, W. H. Fatigue and hours of service of interstate truck drivers. II. Psychomotor reactions. *Publ. Hlth. Bull.*, Washington, 1941, No. 265, 135-177.



## Dimensional Analysis of Motion: VII. Extent and Direction of Manipulative Movements as Factors in Defining Motions<sup>1</sup>

Shelby J. Harris and Karl U. Smith

*University of Wisconsin*

In earlier investigations the problems of extent and direction of travel movements as factors in determining the duration of the manipulative and travel components of motion have been investigated (5, 6). Contrary to assumptions and observations in the fields of human engineering and time and motion study, these studies indicate that greater travel distances increase the duration of both the travel and manipulation components of a motion. The same experiments also show that the direction of travel movement affects only the travel time of the motion. The present experiment extends this line of investigation by studying the effects of varying the extent and direction of manipulation on the component movements of travel and manipulation in the motion pattern.

### Methods and Procedure

The apparatus (Figure 1) used in this study consists of an electronic motion analyzer which has been named the analytic reactometer (3). This device is designed in terms of two main features: (a) control of the space dimensions of the motion pattern; and (b) separate measurement of the manipulative and travel components of motion through the use of special electronic relays (4). The electronic methods of motion analysis, as adapted to the present apparatus, are based on the principle of making the human operator a key in a circuit consisting of the performance situation, the operator, an electronic relay and precision time clocks. When the subject operates one of the switches of the apparatus, he activates a vacuum tube relay causing the manipulation-time clock to run as long as he is in contact with the switch. When he ceases contact with the switch, another relay is thrown, causing the travel-time

clock to run. This clock is stopped and the manipulation-time clock started again as soon as another switch is touched.

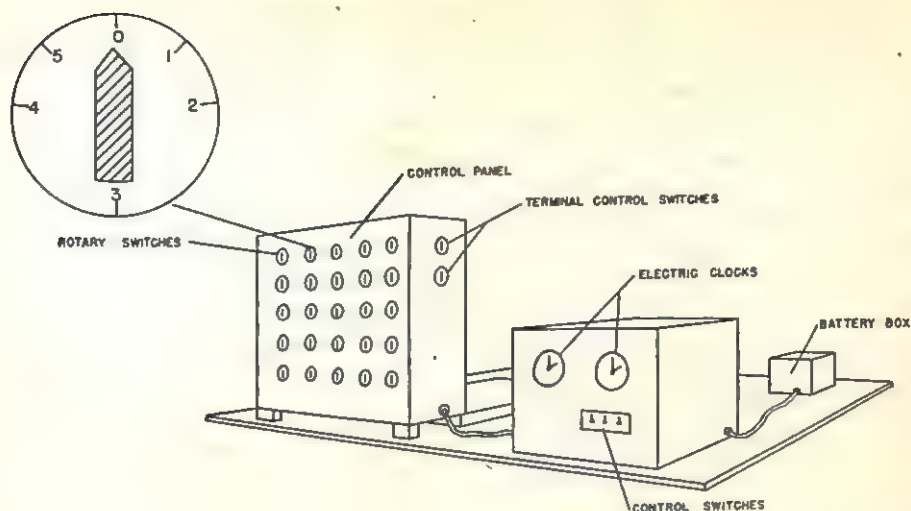
The planned performance situation employed in the experiment consists of a control panel, 45.7 cm. square, on which are mounted 25 rotary switches. These switches are mounted in five rows of five switches and are spaced by a distance of 7.6 cm. Each switch has 17 possible settings spaced at intervals of 20 degrees. Settings of 40, 80, and 120 degrees clockwise and 40, 80, and 120 degrees counterclockwise are marked on the dials.

In the design of the experiment three extents of manipulative movement, 40, 80, and 120 degrees, and two directions of such movement, clockwise and counterclockwise, were used, thus providing a total of six separate conditions. The over-all pattern of travel movement was the same on all tests with the subject starting at the top of the control panel and working from left to right through all five rows of switches. Each subject performed each of the six tests once at approximately the same time on seven successive days. All performances were carried out with the right hand.

A total of 42 right-handed men and women students from the elementary classes in psychology at the University of Wisconsin served as subjects.

In order to control the effects of the sequence and ordinal position of the six experimental conditions a  $6 \times 6$  latin square with seven replications of the same square was used. Subjects were assigned to a given sequence of tests in order of appearance for the experiment and were required to repeat the same sequence on seven successive days. Separate analyses of variance were performed on the travel and manipulation time data for the first and seventh days of the experiment only. Performances on these days are considered to represent unskilled and skilled lev-

<sup>1</sup> This research has been supported by funds voted by the Legislature of the State of Wisconsin, and assigned by the Graduate School Research Committee, The University of Wisconsin.



## A SCHEMATIC DIAGRAM OF THE ANALYTIC REACTOMETER

FIG. 1. Diagram of the analytic reactometer showing the arrangement of controls on the panel and the timing mechanism. The inset illustrates the design of the individual manual control. The 120 degree extents, clockwise and counterclockwise, which were used in the experiment are not shown on the dial.

els of performance. The choice of seven days of practice was arbitrary and, therefore, the term "skilled" is not intended to imply a maximum level of performance.

Learning curves of the component movements for the various tests over the seven days were also constructed.

During the testing procedure the subject was seated on a chair and his height adjusted so that his eye level was approximately equal to the top row of switches. The subject was instructed to move the chair toward or away from the control panel to a comfortable position but required to keep it centered in front of the panel throughout the test session. Prior to each of the individual tests on the first day, the subjects were given a practice trial consisting of turning the first 10 switches to the appropriate position. Each subject was instructed to turn the switches to the appropriate position as rapidly as possible and at the same time to be careful to position the switches accurately. Although error scores were not used in the analysis of the data, they were recorded in order to discourage subjects from becoming careless with regard to accuracy.

## Results

Figure 2 shows the learning curves for travel and manipulation movements for the three conditions involving clockwise direction of manipulation. Analogous curves for counterclockwise direction were obtained, but since the two sets of curves are much the

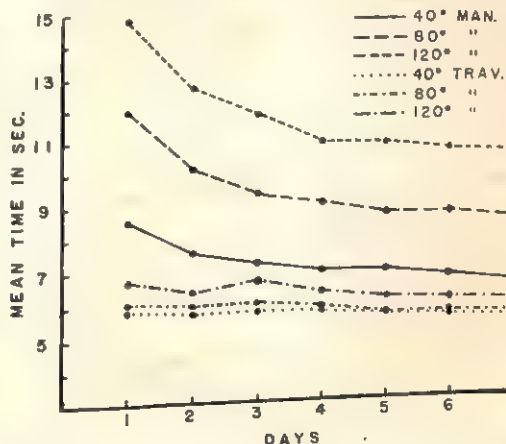


FIG. 2. Learning curves for 40, 80, and 120 degrees extent of manipulation in a clockwise direction. The mean times for 42 subjects are shown separately for the manipulation and travel components of motion. Analogous curves for counterclockwise direction are similar to those shown.

Table 1

Per Cent Change in Manipulation and Travel Time from Day 1 to Day 3, Day 3 to Day 7, and Day 1 to Day 7 under the Different Experimental Conditions

Exp. Cond.	Manipulation			Travel		
	Day 1- Day 3	Day 3- Day 7	Day 1- Day 7	Day 1- Day 3	Day 3- Day 7	Day 1- Day 7
40 Deg. Right	16.08	8.75	23.43	1.20	5.92	7.06
80 Deg. Right	21.11	8.53	27.84	.83	8.00	8.76
120 Deg. Right	20.00	10.14	28.11	1.49	10.41	11.74
40 Deg. Left	16.13	7.09	22.08	-2.33	7.53	5.38
80 Deg. Left	20.23	5.80	24.86	-3.95	10.78	7.25
120 Deg. Left	15.19	6.27	20.50	2.96	11.05	13.68

same, only those for clockwise rotation are shown. It is apparent that over the seven-day period the manipulation-time scores show considerably greater improvement than the travel-time scores. The major difference in the rate of improvement between the two motion components is during the first three days of practice. Quite similar practice effects are found for the two component movements over the last four days of the experiment. These changes in performance are shown in Table 1 in terms of per cent change from day one to day three, from day three to day seven, and from day one to day seven. An analysis of variance performed on the data for days one and seven indicates that the changes in duration of both travel and manipulation movements over the seven-day period are significant at the .001 level of confidence.<sup>2</sup>

The data on the effects of direction and extent of manipulation were first examined for homogeneity of variance between the different experimental conditions. A Bartlett chi-square test for homogeneity of variance applied to the time-score data for days one and seven proved significant, thus necessitating a logarithmic transformation of the data. All of the analyses were performed on the transformed data.

In order to evaluate the effects of the various sequences of tests, analyses of variance

were performed on the travel and manipulation data from the latin squares for days one and seven. In no instance was the sequence of tests a significant variable.

Figures 3 and 4 show the relation between the mean travel and manipulation times and the extent and direction of manipulation on days one and seven respectively. The means shown in the figures have been computed on the transformed scale and then converted back to the original scale. As may be seen from the graphs, the mean manipulation times increase considerably with increased extent of manipulation for both clockwise and counterclockwise directions of movement. The function relating the two is approximately linear. With the exception of the 120-degree movement on day one, the mean manipulation times for clockwise direction are consistently less than the comparable figures for counterclockwise direction. The mean travel times also increase with increased extents of manipulation in both directions, but the increase is not as pronounced as it is for manipulation times. Inspection of the travel time curves suggest that the relation between mean performance and extent of manipulation also approximates linearity.

Summary tables for the analyses of variance performed on the relations discussed above are shown in Table 2. Extent of manipulation is significant at the .05 level or greater for both the travel and manipulation components on both days one and seven. Direction of manipulation is significant only for the manipulation component on day seven.

<sup>2</sup> The raw data and the summaries for the analysis of variance for this experiment are on file at the University of Wisconsin in the master's thesis of Mr. Shelby Harris entitled "Dimensional Analysis of Motion: The Factors of Direction and Extent of Manipulative Movement in Motion."



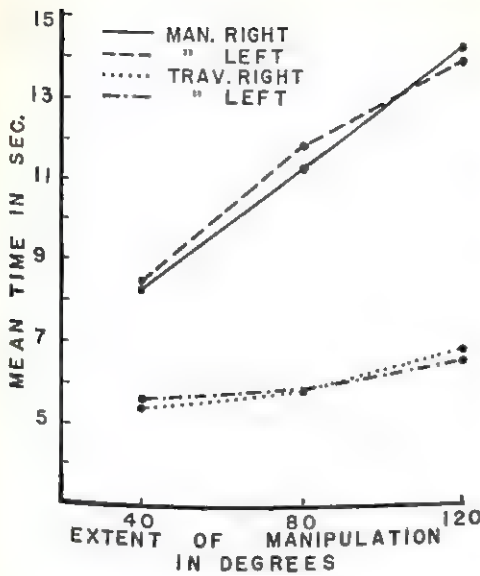


FIG. 3. Mean times for the 42 subjects for 40, 80, and 120 degrees extent of manipulation in clockwise and counterclockwise directions. The means shown are for the first day of the experiment.

Clockwise direction is significantly superior to counterclockwise direction in this instance at the .05 level. Direction does not have a significant effect on the manipulation or travel times on day one, or on the travel component on day seven. None of the direction by extent interactions is significant.

Table 2

Summary of Analysis of Variance for Direction and Extent of Manipulation for Travel and Manipulation Components of Motion

Source of Variation	Manipulation		Travel	
	d.f.	F	d.f.	F
Day One				
Direction	1	—	1	—
Extent	2	72.91***	2	14.07***
Interaction	2	—	2	—
Error	246		246	
Day Seven				
Direction	1	3.04*	1	—
Extent	2	59.19***	2	4.29**
Interaction	2	—	2	—
Error	246		246	

\* Significant at .05 level.

\*\* Significant at .01 level.

\*\*\* Significant at .001 level.

## Summary and Conclusions

Forty-two subjects were tested on a task involving repetitive switch turning under six different experimental conditions. These conditions consist of three extents of manipulation, 40, 80, and 120 degrees, and two directions of manipulative movement, clockwise and counterclockwise. Special devices, involving electronic motion analysis techniques and a special planned work situation, are used to obtain separate measurement of the travel and manipulation components of motion under controlled conditions. Each subject performs one trial under each of the experimental conditions on seven successive days.

Learning curves for the travel and manipulation components of motion are presented. Analyses of variance, performed on the data for days one and seven, are summarized to indicate the significance of differences in the duration of travel and manipulation movements in relation to the direction and extent of manipulation.

The results of the study may be summarized as follows:

1. Manipulation movements show a considerably greater improvement due to practice

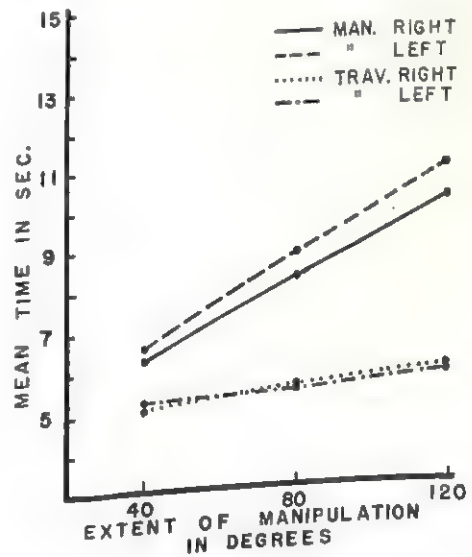


FIG. 4. Mean times for the 42 subjects for 40, 80, and 120 degrees extent of manipulation in clockwise and counterclockwise directions. The means shown are for the seventh day of the experiment.

than travel movements. This differential learning effect is evident primarily over the first three days of practice. The change in performance from day one to day seven is highly significant for both motion components.

2. Duration of both manipulation and travel time is significantly increased with greater extents of manipulative movement and, at least for the extents investigated, the relations are approximately linear.

3. Clockwise direction of manipulative movement is observed to be significantly superior to counterclockwise movements for the manipulation motion component on day seven. Direction of manipulation is not related to duration of manipulation time on day one, nor were travel times affected by the direction of manipulation on day one or day seven.

Previous studies (6) have shown that the durations of both the travel and manipulation movements are related to the distance of travel movement. Inasmuch as the observations reported here show that increasing the extent of manipulation lengthens the duration of both manipulation and travel, it appears that varying the extent of any of the components of motion will have effects on other component movements involved in the pattern. Direction of travel movement has previously been shown to affect only the travel component of the motion (5). Thus, it appears that direction of movement, at least under conditions investigated thus far, influences only the component of motion within which the directional factor occurs. Further research to determine more precisely under what conditions direction becomes a

relevant variable in determining the duration of movement is needed.

The effort to systematize the study of human motions in industry and in some phases of human engineering have led both psychologists and engineers to assume that varying the extent of movement has no influence on the duration of such movement (1, 2). Such assumptions have been proven, through dimensional and component motion analysis, to be erroneous for both manipulative and travel components of human manual motion. Both manipulation and travel times vary significantly with increasing extent of motion. In addition, the increase in extent of one of these component movements in a complex task produces, through interaction of the different movements, significant increase in duration of the other component movements in the task.

Received June 18, 1953.

#### References

1. Barnes, R. M. *Motion and time study* (3rd Ed.). New York: John Wiley, 1949.
2. Ellson, D. G. The application of operational analysis to human motor behavior. *Psychol. Rev.*, 1949, 56, 9-17.
3. Harris, S. J. and Smith, K. U. Dimensional analysis of motion. V. An analytic test of psychomotor ability. *J. appl. Psychol.*, 1953, 37, 136-142.
4. Smith, K. U. and Wehrkamp, R. A. A universal motion analyzer applied to psychomotor performance. *Science*, 1951, 113, 242-244.
5. Von Trebra, Patricia A. and Smith, K. U. Dimensional analysis of motion. IV. Transfer effects and direction of movement. *J. appl. Psychol.*, 1952, 36, 348-353.
6. Wehrkamp, R. and Smith, K. U. Dimensional analysis of motion. II. Travel distance effects. *J. appl. Psychol.*, 1952, 36, 201-206.

## Discussion of Gilliland and Newman's "The Humm-Wadsworth Temperament Scale as an Indicator of the 'Problem' Employee"<sup>1</sup>

D. G. Humm and Kathryn A. Humm

*Humm Personnel Consultants, Los Angeles, California*

The first question to be raised is that of the title of the article in question. The methodology reported does not represent the methodology recommended for using the Humm-Wadsworth Temperament Scale in the appraisal of employees and applicants for employment. If the article had been given some such title as "The Integration Index and Component Control Measures Computed from the Humm-Wadsworth Temperament Scale as Indicators of the 'Problem' Employee," we would be less dissatisfied with it.

It is implied by Gilliland and Newman that they used Humm's procedures in classifying their subjects according to "risk," as described in our study of Los Angeles policemen<sup>2</sup> and as discussed in personal conference.

On the contrary, we recommend evaluating Humm-Wadsworth findings for each subject tested by considering each of the following: (1) the raw scores, corrected for response-bias, in comparison with the scores of the subjects of the original standardization study;<sup>3</sup> (2) the degree of response-bias itself, since atypical response-bias has been found to be an indicator of tendencies to problem behavior; (3) the positions of the seven components in the distributions of the scores of employed subjects, but without any implication that conformity to the central tendency is necessarily desirable; (4) the relationship of

the Normal component to each of the other components (the component control measures) and to the temperamental pattern as a whole (the integration index),<sup>4</sup> and (5) finally, the temperamental pattern as a whole, derived from all of the measures previously mentioned, and indicating which components are likely to be conspicuously manifested in the subject's behavior and whether their manifestations will be desirable or undesirable in the situation for which the subject is being considered.

In general personnel work, we assign a risk rating on the basis of Humm-Wadsworth findings alone only when those findings are so unmistakably unfavorable as to constitute an insurmountable handicap even if all findings concerning ability should be found to be favorable. In our report of the study of Los Angeles policemen, we attempted to make it clear that we used the Humm-Wadsworth findings alone, without partialling out other factors related to job success, because the policemen in question had been pre-selected by the civil service procedure.

When Gilliland and Newman classified their subjects on a scale which they do not identify and which we cannot recognize for the Integration Index and the Component Control Measures and then assigned risk ratings on the basis of Very Good for all ratings above five and Very Poor whenever any two ratings were as low as one, they were using a procedure we have never recommended and strongly disapprove.

The explanations offered by Gilliland and Newman for the outcome of their study seem to us not to follow from the data reported, i.e.; "(1) the test may not adequately meas-

<sup>4</sup> Humm, D. G., and Humm, Kathryn A. Measures of mental health from the Humm-Wadsworth Temperament Scale. *Amer. J. Psychiat.*, 1950, 107, 6, 442-449.

<sup>1</sup> Gilliland, A. R. and Newman, S. E. The Humm-Wadsworth Temperament Scale as an indicator of the "problem" employee. *J. appl. Psychol.*, 1953, 37, 176-177.

<sup>2</sup> Humm, D. G. and Humm, Kathryn A. Humm-Wadsworth Temperament Scale appraisals compared with criteria of job success in the Los Angeles Police Department. *J. Psychol.*, 1950, 30, 63-75.

<sup>3</sup> There were seven pairs of groups, rather than seven groups, and they were not "relatively pure types." A partial regression technique had to be used to "purify" the data. See: Humm, D. G. and Wadsworth, G. W., Jr. The Humm-Wadsworth Temperament Scale. *Amer. J. Psychiat.*, 1935, 92, 1, 163-200.



ure the components it purports to measure"—no data are presented to indicate whether or not the behavior of the subjects differed from the behavior that might have been predicted from the Humm-Wadsworth results; "(2) these components may not be essential to success in this industry"—the study investigated only a specific set of measures and did not do this in a way which could justify such a conclusion; "(3) the company cannot distinguish between satisfactory and unsatisfactory workers"—no data are presented of

the procedures used by the company for determining satisfactory work or for deciding to discharge an employee.

The only conclusion we are able to draw from this study is that it supports our own contention that over-simplified procedures are inadequate for appraising workers, but that it offers no evidence as to the effectiveness of the Humm-Wadsworth, properly used, as one of the tools for personnel appraisal.

*Received July 31, 1953.*

*Published out-of-turn by the editor.*

## Applied Psychology in Action

### Comment on Word Meaning

Fred L. Wells

Department of Hygiene, Harvard University

In the December 1953 issue of the *Journal of Applied Psychology*, Dr. H. D. Hadley has an insightful note: "The Non-Directive Approach in Advertising Appeals." For the text, I wonder if there would be any interest in a distinction between *credibility* and *credulity*. The latter term seems the one fitting Dr.

Hadley's context better, but the word used is *credibility* (page 496, line 5 from end, page 497, bottom of column one). There is apparently an "obsolete" use of *credibility* in Dr. Hadley's sense. (See Webster.) If interpreted in the current usage, this makes the author's meaning difficult to follow.

### A Note on "The Non-Directive Approach in Advertising Appeals"<sup>1</sup>

Mary Epstein

George Peabody College for Teachers

The December 1953 issue of *J. appl. Psychol.* carried an article, the aim of which was to point up some similarities between certain types of therapy and advertising techniques. One of the conclusions reached, that "the non-directive technique is quite comparable to the inferred technique in advertising . . .," needs further elaboration.

Showing the benefits of a product, without intention to sell (the inferred technique) seems, according to some standards, superior to the direct appeal, where the advertiser tells the consumer to buy the product. Often, the direct appeal involves threat to the buyer's values, particularly when his attention is called to the fact that the purchasing of another product than the one advertised may lead to various undesirable results. The inferred technique minimizes the effects of threat, by emphasizing the acceptability of the "thing," and by associating it "with very acceptable things, persons or events."

Comparison between non-directive therapy<sup>2</sup> and inferred advertising can possibly be made

concerning the attempt to reduce threat. The advisability of reducing threat, in any field of human endeavor, is psychologically sound. Further comparison, however, can only be drawn by doing injustice to the basic principles of client-centered therapy.

A closer examination of the assumptions on which non-directive therapy rests reveals that a belief in the client's ability to develop his own value system is a *sine qua non* of successful therapy. To facilitate this process, the therapist tries to minimize, as much as possible, the effects his own value system might have on that of the client.

The principle of non-interference is not applicable in the field of advertising, because, carried to its logical conclusion, it would mean that the advertiser not sell at all. Placed in the non-directive framework, where selling the client on anything is *verboten*, the advertiser would be in no better position to make a product look favorable in the eyes of the buyer than is the client-centered therapist in the position to steer the client's value judgment in the direction of his own. There is a fundamental difference between non-directive therapy and advertising. The difference lies in the realm of commitments and intentions. The advertiser is committed to and intends

<sup>1</sup> Hadley, H. D. The non-directive approach in advertising appeals. *J. appl. Psychol.*, 1953, 37, 496-498.

<sup>2</sup> Rogers, C. R. *Client-centered therapy*. Chicago: Houghton-Mifflin Co., 1951.

to sell. The therapist aims to help the client achieve more satisfactory adjustment, i.e., happiness, regardless of the values adopted or discarded in the process of therapy.

Whereas there is no reason to doubt that

some of the elements of non-directive therapy, such as reduction of threat, might prove helpful in raising advertising standards, an unqualified comparison between this type of therapy and advertising is inappropriate.

THE JOURNAL OF APPLIED PSYCHOLOGY  
Vol. 38, No. 2, 1954

## The Measurement of Academic Freedom

Willard Kerr

*Illinois Institute of Technology*

Can academic freedom be measured? Through most of man's history it has existed in such minute quantity and excited so little interest as to discourage evaluation. Today, despite the current wave of anti-intellectualism, academic freedom exists in a magnitude unknown to antiquity. But from one institution to another, where scholars work, there is great variation in academic freedom.

In 1953, the Academic Freedom Committee, Chicago Division of the American Civil Liberties Union, attempted to measure academic freedom in each of the more than 50 institutions of higher learning in the State of Illinois. With the aid of other members of the committee and the ACLU booklet entitled *Academic Freedom and Academic Responsibility*, a two-page "test" of academic freedom was constructed. It was called the "Academic Freedom Survey." It contained twelve items on rights of students, seven on rights of teachers, and four general rights. Each item was answered on a three-point scale of "Extent to which right is effectively assured—complete; as a general rule; very little or none." Possible scores could range between 23 and 69.

**Design.** Approximately 200 of the questionnaires were mailed to Illinois colleges addressed to: (a) one administrator, usually the president; (b) one or more professors; and (c) one or more student leaders, usually the newspaper editor or student council president.

**Results.** A total of 73 replies was received, and, while analysis of the data still continues, the obtained data do indicate substantial freedom variations. The most entrenched freedoms are: *for faculty*, freedom from spe-

cial requirements (oaths), of association in faculty organizations, of citizenship activities, and of research; and *for students*, freedom of choice of faculty advisers. For faculty, the least secure freedoms relate to faculty self-government, to tenure (security), and freedom to criticize curriculum and administration. For students, the least secure are to hear outside speakers, to criticize faculty and administration, to organize associations and affiliate nationally, of press, of petition, and of reasonable off-campus activity.

These results suggest that serious deficiencies exist in academic freedom for both faculty and students, particularly the latter. While our young people are expected to hold themselves ready to fight and die for our country, yet we withhold from them the reasonable freedoms which make for individual responsibility and character growth. This statement is qualified by the fact that most institutions maintain an admirable situation with respect to most of the freedoms studied. Total results as analyzed to date indicate that the "test" may be most useful as a diagnostic instrument to indicate areas for remedial attention in a given individual institution.

The Chicago Division, ACLU Academic Freedom Committee now plans to resume such surveying, but this time with "Student," "Administrator," or "Faculty" stamped on each form in order to establish points of agreement and disagreement among these three groups and thus more clearly delineate the freedom areas requiring attention in each institution. This kind of impartial service in the interests of human freedom is traditional in the history of the American Civil Liberties Union.



## Book Reviews

Lincoln, J. F. *Incentive management*. Cleveland: The Lincoln Electric Company, 1951. Pp. 280. \$1.00.

This volume is written by the president of the Lincoln Electric Company, which manufactures electric welding equipment. It is an exposition of the rationale for the system of incentive management upon which the company is run. The rationale, as digested by the reviewer, is as follows:

1. The primary goal of industry is to make a better product to be sold to more people at a lower price; a reasonable profit to the stockholders is also important but should be a secondary by-product.

2. This goal is possible only under conditions of free enterprise and ever increasing efficiency of operation.

3. Such levels of efficiency are possible only if workers are motivated to develop their latent abilities, which are limitless under proper incentive conditions.

4. Workers will develop their latent abilities only if they are given a direct reward for their individual contribution to production.

5. This direct reward is obtained through an incentive wage system and recognition of the individual's ability.

As can be seen from the above, the rationale is essentially an application, in modern industry, of the law of competitive struggle for existence and the survival of the fit. Fired by the knowledge of reward and recognition for demonstrated superiority, human beings have limitless possibilities of improvement. Through "intelligent selfishness" man strives on and on toward perfection and great strides of progress result.

The author is thoroughly convinced that this rationale is true. Why is he so sure? Because under his management according to these beliefs, his company has become the most productive organization in the industry. Charts and tables (in the appendix) show that prices of Lincoln-made products steadily declined from 1933 to 1949 while those of comparable products increased. Sales value per employee is double that of the average in other industries and other companies in the same industry. There is no union; there were no work-stoppages due to labor-management disputes in any year from 1934 to

1949. Productivity increased 15% per year from 1934 to 1949 compared with only 3% per year for all manufacturing industries. The average Total Compensation per Employee was \$7701 in 1950 compared with between \$3000 and \$4000 for six other well-known companies, some of which are competitors. As the author states, "The conclusion that must be drawn from these facts is obvious . . . The American economy must adopt incentive management."

This is a very difficult book for a psychologist to evaluate. Research-oriented, he looks for a statement of hypotheses, description of procedures designed to test the hypotheses, presentation and interpretation of results, and conclusions derived from the findings. In this book, however, the author presents merely an exposition of the "hypotheses," which he presents as axioms, and his proof of their validity is the ultimate criterion of the production record of the company. Just how the principles are translated into operation procedures and what the relative contribution of these procedures is to the overall success of the company are not given, either in this volume or in a previous book entitled *Lincoln's Incentive System*. The reader is left with a feeling of something missing and with nothing to evaluate objectively and critically. It is like a father who dogmatically states that proper diet results in healthy children and proudly points to his six-foot son as the proof. Can one conclude that diet increases height? Certainly not without knowledge of what diet how administered, and of related variables such as exercise, height of parents and grandparents, etc.

This is not necessarily to deny his thesis; it is just that he hasn't proved it. In fact, except for the fundamental weakness of any explanation of human behavior which rests upon a single source of motivation, the exposition of his thesis is a fairly well-reasoned, consistent, thought-provoking presentation. It is not difficult to accept his premises of the desirability of direct and immediate reward for individual effort, of the importance of overt recognition of individual achievement, of individual identification with group goals through stock ownership in the company, of the greater value of the earned security re-

sulting from self-confidence and assurance of reward than that granted by a paternalistic employer or government. The author has a supreme confidence in the ability of man to rise to new heights of performance and emphasizes employee development rather than selection. The industrial progress which has made us a leading nation can be maintained only with the continuous change resulting from the struggle for existence under conditions of free competition. The "profit motive" is reinstated in full force but with the "profit" more equitably distributed—the major share going to the consumer, through lower prices, and thereby to the workers who themselves are consumers.

One wonders, however, whether the Lincoln Incentive System is universally applicable. It is conceivable that it works at Lincoln primarily because it is unique. Nowhere in the book does Lincoln discuss selection standards or turnover or what happens to those employees who do not produce at a high level. Assuming that they are weeded out or allowed to weed themselves out through low returns from piece-work wages, the company may have a highly selected work force—selected in terms of their responsiveness to the particular type of incentive and the level of performance required by his system. One is reminded of the tremendous spurt to production accompanying Ford's introduction of the \$5 a day wage but also of the levelling off which resulted as time went on.

One is forced to admire the positiveness with which Lincoln obviously believes in the philosophy underlying his system and the courage with which he applies it. He is not "afraid" to pay a low-level production worker whatever the worker can earn under piece-work rates rigidly maintained. He is much more concerned with the benefit to the consumer than to the stockholder, who adds nothing to the productive effort. He has sincerely attempted to put his beliefs into actual practice without compromise and is thoroughly convinced of the validity of his beliefs. As he states in the companion volume previously referred to, "Whatever the conclusions of the reader, there is no doubt that the incentive-management philosophy

outlined herein is fundamental to man, whether he is playing a game, raising a garden, or living a life." I wish that psychologists could be as positive in their knowledge of human behavior and its application to life situations.

Psychologists can profitably read this book. They will feel successively annoyed, amused, disturbed, provoked, and challenged. The price is only one dollar, an example of the author's philosophy of lower prices for the consumer.

Albert S. Thompson

Teachers College,  
Columbia University

Viteles, Morris S. *Motivation and morale in industry*. New York: Norton, 1953. Pp. xvi + 510. \$9.50.

Any new book by Viteles is bound to command attention. As one of the first and still one of the leading industrial psychologists, his work merits and gets the attention of workers in the field.

Viteles' well known *Industrial Psychology* was first published in 1932, and since that time has been considered a classic, if not the classic text in the field. Drawing heavily upon the experience of psychologists in the laboratory as well as in industry, Viteles gave a comprehensive picture of the development and current status of industrial psychology which at that time was considerably less robust than it is today. Advocating the view that the scope of industrial psychology was as extensive as that of psychology itself, Viteles nevertheless emphasized individual differences.

In many respects, *Motivation and Morale in Industry* is a continuation of *Industrial Psychology*. To a considerable extent Viteles has repeated his earlier pattern, but with a shift in emphasis from the individual to the group. He is still interested in increasing productivity, but his frame of reference is employee satisfaction and industrial harmony. Again he has drawn heavily upon the experience of psychologists in the laboratory as well as in industry. Again he has synthesized the work of other persons. Again he has pointed out trends and probable trends.

*Motivation and Morale in Industry* is divided into five parts. The first, consisting of



three chapters, is introductory in nature. It deals primarily with the economic man and the inadequacy of the concept that man can live by bread alone. The fifth part, consisting of four chapters, summarizes and draws together the remainder of the book as well as makes applications and recommendations. The remaining three parts, totaling sixteen chapters, comprise the bulk of the book. They deal with motivational theory, experimental studies, and employee attitude surveys.

*Motivation and Morale in Industry* is eclectic. The bibliography refers to books and articles from all fields of psychology. Psychoanalytical, topological, and Gestalt psychology are represented as well as the more traditional fields. Various allied disciplines are also represented such as philosophy, economics, sociology, endocrinology, medicine, and anthropology. Reference is made to publications in various languages and to research conducted in various countries. Included are Canada, England, Germany, Russia, and the Netherlands. An extensive period of time is covered, from William James to the present. Business and trade publications are referred to; for example, National Industrial Conference Board, National Association of Manufacturers, Factory, and Dun's Review. Nontechnical journals and books are also included such as New York Times Magazine, Survey Graphic, Fortune, and Readers Digest.

The book is scholarly. In fact it is probably too scholarly to be of maximum value and use to the audience to whom it is directed: management in business and industry. The book is far more suitable for students preparing for work in management. The typical business executive is impeded rather than helped by phrases such as "sine qua non," "in vacuo," and the like. He does not need nor desire references in foreign languages. He does not readily accept the involved sentence structure or the laborious style used by Viteles. A Flesch count of this book would place it far beyond the "comfortable" reading level of most management people. A re-write of *Motivation and Morale in Industry* will be required if it is to gain wide acceptance in industry. Viteles did this earlier when *Industrial Psychology* was re-

written to fill the need for a shorter and simpler volume, and was published in 1934 under the title *The Science of Work*.

Viteles has written his book from a theoretical and experimental viewpoint. This is well illustrated by his statement "Effective results can be achieved only through systematic research conducted within a sound theoretical context" (p. 66). Would that all workers in industrial psychology took this view!

In a sense this book is too much a book of readings in motivation and morale in industry. Many of the studies are weak, but Viteles has done an excellent service in collecting these studies in such way as to illustrate the primitive status of the field. Frequently he has added his penetrating insights relative to such studies. Nevertheless, the reviewer regretted that Viteles had not taken a more directly critical view. It were as though a skilled surgeon held his scalpel to the skin but neglected to make a sharp and deep incision. Why? Does Viteles feel less sure of himself in the area of the group than in that of the individual? Do his own feelings emphasize the individual, but his intellect tell him to emphasize the group? Or, is he a highly tolerant man who is convinced that more harm than good would result from a more critical attitude at this time?

The book was published prematurely in one respect. References were added after the type was set without changing numbering. Thus, the same number frequently appears successively, the second being followed by a letter subscript. This may be minor, but is apt to give some readers the impression of haste or carelessness which is inappropriate in a book of this type. It is hoped that reprinting will correct this defect.

In spite of its deficiencies, this is a book which should be studied carefully by all who profess to be interested in industrial psychology. It pulls together much material which has lacked structuralization. In so doing Viteles has done a valuable service albeit the material is primitive. Out of such syntheses can come considerable improvement in future motivational theory and experimentation.

In writing this book, Viteles has not blindly jumped onto the band wagon of "group



think." This is particularly refreshing inasmuch as so many psychologists seem to disregard their own teachings and to follow the "all or none" hypothesis in evaluating schools, viewpoints, methods, and procedures in the field of psychology. As Viteles points out, "The emergence of a 'social psychology' does not require or justify the abandonment of 'individual psychology' in approaching or solving the problems of motivation and morale in industry" (p. 391).

Clifford E. Jurgensen

Minneapolis Gas Company

Redfield, Charles E. *Communication in management*. Chicago: University of Chicago Press, 1953. Pp. xvi + 290. \$3.75.

Redfield's book presents an excellent broad view of the problem of communication in industry as well as information on how to handle rather specific problems. The author states that while the means of communication have now reached their greatest development, "intelligibility" in industrial communication is at its lowest stage in history. This, he says, is due to: (a) the increasing size of modern organizations; (b) lack of training in wise language usage; and (c) the specialization and segmentation of work today. Of the importance of communication, however, Redfield leaves no doubt when he quotes *Fortune's* new motto for business,—"Communicate or Founder."

The book is arranged in five parts. The first part provides a general introduction to the problem, and contains highly useful guiding principles for effective communication. It is necessarily general in scope, but it does seem to give too little attention to one aspect of communication, effectiveness as a function of the educational differences of "communicator" and "communicatee." The goal is stated as having members of the audience improve their language facility (as well as having the communicator improve his way of using language). This is desirable, but the practical question remains whether or not communication *can* be effective if the reader or hearer cannot understand. It would have seemed worth while to present more information on how to make writing and reading more understandable, and how to check on this through readability formulas. Redfield, it

should be said, does not deny the importance of the problem, however, for he says earlier that "In the America of the 1950's, literacy will have to be measured in terms of comprehension of transmitted ideas and concepts."

Part II of the book takes up "communication downward and outward," the most important aspect of which is order-giving. After a description of kinds of orders, Redfield goes on to a discussion of oral versus written presentation. He then takes up individual messages and circulars, manuals, and handbooks. The presentation is thorough, but this very thoroughness in itself leads to some generalizations that may not always be accurate. For example, Redfield says a safe rule of thumb in distinguishing manuals and handbooks is that "if personal pronouns appear in the text, it is a handbook and not a manual." This is, however, a minor point as far as the whole presentation is concerned.

In Part III, Redfield presents "communication upward and inward." He gives chief attention to "administrative reporting" as essential to the executive, but also takes up suggestion (and complaint) systems, interviews, and employee opinion polls. The overall presentation is excellent, and should introduce new approaches to many readers.

Part IV of the book is an interesting presentation of "horizontal communication," or such cross-talk as clearance, review, and conferences. Horizontal communication, as Redfield points out, is of increasing importance because of growing specialization in industry.

In the final section of the book (Part V), Redfield presents his views of the future of communication in management. The presentation is largely in terms of organization in management and its relation to communication. Recent changes in organizational structure (reduction of number of management levels) in several large corporations, and the effect on communication, provide interesting reading.

All in all, the book is a valuable one, chiefly for its survey of the field and its complete list of references and selected readings. It should prove useful to most readers concerned with management, but particularly to those who have not recognized the extent of communication that goes on in industry, or

how more effective communication can improve industrial efficiency.

George Klare

*University of Illinois*

Tyler, Leona E. *The work of the counselor*. New York: Appleton-Century-Crofts, 1953. Pp. 323. \$3.00.

During the past four years an unusual number of textbooks on counseling have appeared. Some of the texts have been elaborations or developments of the nondirective point of view; some have been restatements of older points of view modified to incorporate a greater emphasis on counseling as contrasted with diagnosis; and one or two have been attempts at something like a synthesis. Tyler's text belongs in this last category, and, in this reviewer's judgment, is outstandingly successful in this class.

As the title indicates, Tyler has attempted, not to describe a theory of counseling, but to write of the peculiar work of the counselor, marshalling ideas from experience and from research to throw light on how counseling may most successfully be done. It is therefore an eclectic book in its approach, predominantly nondirective in its philosophy and techniques, but making use of the contributions of testing, occupational information, and environmental resources in a manner more commonly associated with other points of view. Tyler makes her own synthesis of these approaches. The result is a very readable text, suitable for relatively unsophisticated students, in which each chapter concludes with a concise critical summary of relevant research which makes the text appropriate for students with more background and for practitioners.

The functions of the counselor in modern society are effectively dealt with in Chapter I, thus starting out by putting the counselor's work in good social and psychological perspective. Chapter II discusses interviewing, stressing the perceptual skills of the counselor and reflection of feeling as a tool but pointing out that these are procedures used by a warm person communicating with another, not tricks of the trade. Nondirective theory is revised here, for example, with the recognition that verbal structuring is of little value, that effective structuring is behavioral

rather than verbal. Chapter III deals with records in a manner that is refreshing among texts of this type: instead of discussing the construction of cumulative records, Tyler treats them as aids to counseling, as sources of hypotheses to explore in counseling, as a means of orientation to a client rather than as bases for diagnosis. She conceives of the counselor's province as being the client's feelings and attitudes, not objective facts, and she would leave these and the manipulation of the environment largely to other personnel workers.

The chapter on diagnosis therefore recommends that counseling not be organized around this activity, as it typically is in non-Rogerian settings, but that diagnostic activities be relied upon for initial screening and particularly as means of helping the client to understand himself. Data showing that clinical predictions are not valid, but that counseling with tests improves vocational decision-making are cited, and ways of helping clients use test results are discussed in a manner which effectively brings together the contributions of nondirective and diagnostic counseling. Chapters V and VI deal with tests, leaving data on the construction and validation of specific tests to other textbooks, and concentrating on what tests can contribute to the self-understanding of the client and how the counselor can use them for this purpose. The generally admitted desirability of at least one nondirective interview before testing so that problems may be aired, the more debatable advantage of testing by batteries instead of giving a single test when interviewing brings up the need for that kind of fact and other tests as other facts are needed, and the equally debatable value (in this reviewer's opinion) of written reports for clients, are brought out.

The chapter on occupational information also stresses the use of such information in counseling, although brief attention is paid to sources in passing. Thus the distinctive emphasis of this text is maintained, relying on standard texts for information on sources and tools and concentrating on how the counselor uses them in counseling. The stress on occupational information which characterized early vocational guidance, the later rejection of this method by some in favor of testing



and still later by others in favor of counseling concerning attitudes, are placed in nice perspective (although some details of historical explanation are incorrect as in the failure to recognize that early writers such as Parsons also advocated self-understanding and counseling), and a synthesis of these approaches and methods such as that which characterizes much of the best contemporary counseling is achieved. Occupational information is seen as a means of reality testing.

Chapter VIII deals with psychotherapy, and Chapter IX with decision-making interviews, thereby putting this text practically in a class by itself for comprehensiveness and balance in coverage. Tyler stresses the unity of the person and hence of counseling, laments false distinctions between personal and vocational counseling (still incorrectly attributed to the Veterans Administration), argues in favor of counseling which deals with vocational choice as part of the development of the person, and at the same time recognizes that people do have to make occupational decisions. In dealing with psychotherapy she stresses the importance of the relationship, and makes the nice point that reflection of feeling is not so much a technique of treatment as a means of conveying to the client that communication is taking place. Tyler is appropriately modest concerning our knowledge of psychotherapy, and points out issues concerning which we lack information. The analysis of the processes of decision-making and of counseling in this connection is original and helpful.

In Chapter IX the school counselor is placed in the context of the school as one personnel worker, with the peculiar function of trying *not* to decide things for the student. This is a helpful distinction between counseling and administrative functions, but not one which fits the school counselor's job as structured in most schools, where the counselor is also expected to handle discipline, programming, and a variety of decision-forcing, as contrasted with facilitating, functions. The use of community resources and agencies by the counselor is discussed, but not in any detail.

A chapter on the selection and training of counselors, and one on evaluation, bring the

book to a close. The former mentions various professional associations, but makes no mention of the American Personnel and Guidance Association as that which, in 1950, resulted from the unification of all but one of those listed, follows the style of the Michigan Conference in referring to counselor-psychologists instead of the more recent officially adopted APA term of counseling psychologists, and fails to recognize that many counselors and counseling psychologists are employed in community agencies, hospitals, and industrial or business concerns. It is otherwise up-to-date and helpful, particularly in its discussion of the self-selective functioning of a good counselor-training program provided there has been initial screening for academic ability.

The final chapter is an excellent critical review of evaluative studies, except for the curious failure to note the inadequacy of Latham's study which results from its attempt to relate test scores to occupational success after one year of work (shown by career pattern research and longer-term follow-ups to be too brief and early a period to be meaningful), the even stranger omission of Strong's studies on the occupational predictive value of tests, and the final erroneous conclusion which Tyler therefore reaches, to the effect that tests have no predictive value for occupational success and satisfaction.

Three appendices include an intake form, notes on some interviews, and selected readings. The first two are not coordinated with the text and hence have little value beyond what the reader can derive from examining them himself; the last contains a number of helpful references, but excludes all treatises on counseling other than the nondirective (e.g., Robinson, Hahn and McLean, Williamson), surely a mistake in a text which does as good a job of synthesizing viewpoints as does this.

A few criticisms of details and a few major weaknesses should be mentioned, before reaching an over-all evaluation.

The apparent desire to write a smoothly reading, easily digested text occasionally results in less specificity of facts than is desirable, as in the failure to mention the GATB as the USES test under discussion on page



130. It results, furthermore, in slighting the originators of ideas, for while Roethlisberger and Dickson are mentioned as having developed a nondirective approach simultaneously with Rogers (but in 1939 rather than 1937 as stated), Otto Rank and Jessie Taft are not mentioned as important precursors. Many other points made and ideas expressed in the text appear as though they were Tyler's, with no indication of when they are original, when they are a part of the thinking of contemporary counseling psychologists, or when they are novel ideas first expressed by other psychologists in the literature on counseling. The contributions of others to Tyler's thinking get recognition only if they are research contributions, for the only theoretical contributions acknowledged are the two nondirective sources mentioned above and Bordin and Bixler's article on test selection by clients, although others are clearly traceable in Tyler's writing. Finally, the book should have a subtitle, "In Educational Settings," for as pointed out above it is written in terms primarily of the counselor in a college or university, and to a lesser extent the high school counselor, and disregards the fact that many counselors work in social agency, medical, and industrial settings. This limitation does not lessen the value of the book for theory or technique, but it does make its discussion of some operating problems less valuable than it might be to these other counselors.

In this reviewer's judgment, this book is the first genuine attempt at a synthesis of what we know about counseling. Its predecessors have described approaches developed by individuals or groups of psychologists working together in one setting, and hence have been biased by the theoretical predilections and experimental limitations of the contributors. Tyler has, as pointed out above, a theoretical bias in favor of nondirective counseling, one which caused her inadequately to summarize and evaluate the research on the occupational predictive value of tests. She has apparently had limited experience in other than educational settings, which results in the slighting of community resources and work in other settings. But she has drawn on research and theory regardless of school and has critically examined her own work, and

has thus achieved a breadth of viewpoint, variety of technique, and comprehensiveness of scope which make her book unique. To put it in a nutshell, although it seems to this reviewer that the book shows an as yet incomplete recovery from the impact of the nondirectivists, it is an extremely valuable text which many of us active counseling psychologists would be glad to have written ourselves!

Donald E. Super

*Teachers College, Columbia University*

*Recommended practice for residence lighting.*  
New York: Illuminating Engineering Society, 1953. Pp. 44. \$1.00.

This pamphlet, prepared by the Committee on Residence Lighting of the I.E.S., contains information useful to architects and to applied psychologists who are concerned with specifying illumination which will provide an attractive living space as well as comfortable and efficient vision in the home. Important developments in the field provide a basis for marked improvement over the first *Recommended Practice of Home Lighting* which appeared in 1945. The present pamphlet is concerned mainly with basic lighting requirements for family activities which involve close vision.

It is gratifying to find a strong emphasis placed upon those factors which promote comfortable vision. Among these are: (1) a coordination of decoration (painting) with lighting to achieve satisfactory distribution of illumination; (2) the maintaining of satisfactory brightness ratios in the field of view and the surroundings; (3) selection of light sources; and (4) lighting for specific visual tasks such as sewing, dining, etc.

The numerous pictures and figures illustrating types of fixtures and desirable arrangements of lighting for specific seeing tasks are well chosen. Limitations as well as uses are incorporated into much of the discussion. Helpful materials are given in the appendix: detailed description of typical incandescent lamps and of fluorescent tubes, luminaire classification, lighting maintenance, and glossary of technical terms.

There have been two rather marked increases in the light intensities recommended

in 1953 in comparison with those recommended in 1945: For sewing dark fabrics, 150 from 100 footcandles; average sewing, 80 from 40 footcandles. In several instances the recommended intensities are higher than can be justified by research findings. Except where casual seeing is involved, the tendency is to recommend at least 40 footcandles.

This is, in general, an excellent pamphlet on home lighting. The careful reader with a knowledge of the field can approve of all the material except the recommended light intensities.

Miles A. Tinker

*University of Minnesota*

Bullock, Robert P. *Social factors related to job satisfaction, a technique for the measurement of job satisfaction*. Research Monograph Number 70. Columbus, Ohio: Bureau of Business Research, 1952. Pp. 105. \$2.00.

This monograph is the report of a research study designed to discover the relationship of certain social factors to job-satisfaction and to employ these factors in a scale for the measurement of job-satisfaction. The basic assumption underlying the study is that the individual's work behavior and adjustment depend upon his sentiments and attitudes. It is further assumed that these sentiments and attitudes are a result of his attempt to achieve personal adjustment within at least three separate, interacting social systems: the informal work group, the formal work organization and the larger social community within which the employing industry is located. In this study, job-satisfaction is considered to be an attitude resulting "from a balancing and summation of many specific likes and dislikes experienced in connection with the job."

Two measuring instruments were prepared, a Job-Satisfaction Scale for use as a criterion and a Social-Factor Questionnaire. The Job-Satisfaction Scale was of the multiple answer type patterned closely after the Hoppock scales. The Social-Factor Questionnaire consisted of 129 items designed to inventory conditions on the job, in the home, in the community and attitudes of the worker. Seventy-five of these were in Y, ?, N format, thirty

were Agree, ?, Disagree items. Twenty-four were multiple answer questions sampling personal background information. (All are presented for the reader's examination in appendices to the monograph.)

The instruments were pre-tested on a group of 53 male juniors and seniors in college, all of whom had held full time jobs. Validation was accomplished on this group and on two samples from an animal registration association. One hundred currently employed persons comprised the first sample and 124 ex-employees the second. The Job-Satisfaction Scale was checked by testing its ability to differentiate between groups judged "satisfied" and those judged "dissatisfied" (judgments made by a panel on the basis of personnel data), between individuals who gave "satisfied" answers to three factual questions from those who did not and between current and ex-employees. This last differentiation was required on the assumption that dissatisfaction might be more intense and more frequently associated with termination of employment.

To assess the validities of the Social Factor Questionnaire items, individuals in all three samples were divided into extreme groups on the basis of Job-Satisfaction Scale scores. Each item was then evaluated in terms of the CR of the difference between "Satisfied" group responses and "Dissatisfied" group responses. Objections to the instability of CR in small samples were met by requiring high CR's in all three samples.

The author deserves commendation for his adaptation of the Social Factor Questionnaire to the measurement of job-satisfaction and for his attempt to validate his instruments. All too frequently measures in this area are offered to the public without any systematic attempt at validation. Further research, however, is necessary before these results are generalized to other industries. The population utilized was probably well chosen for the purposes of a prototype study. However, the unusualness of the occupation, the small numbers involved and the fact that its workers were non-union suggest the need for cross-validation.

Howard L. Roy

*Personnel Research Branch,  
TAGO, Department of the Army*



## New Books, Monographs, and Pamphlets

Books, monographs, and pamphlets for listing and possible review should be sent to Donald G. Paterson, Editor, Department of Psychology, University of Minnesota, Minneapolis 14, Minnesota.

- Class, status and power.* Reinhard Bendix and Seymour Martin Lipset, Editors. Glencoe, Ill.: The Free Press, 1953. Pp. 732. \$7.50.
- Science and man's behavior.* Trigant Burrow. New York: Philosophical Library, 1953. Pp. 564. \$6.00.
- The teaching-learning process.* Nathaniel Cantor. New York: The Dryden Press, 1953. Pp. 350. \$2.90.
- Fundamental psychiatry.* John R. Cavanagh and James B. McGoldrick. Milwaukee: The Bruce Publishing Company, 1953. Pp. 582. \$5.50.
- The transfer value of guided learning.* Robert C. Craig. New York: Bureau of Publications, Teachers College, Columbia University, 1953. Pp. 85. \$2.75.
- The role of growth hormone in carbohydrate metabolism.* R. C. De Bodo and M. W. Sinkoff. New York: The New York Academy of Sciences, 1953. Pp. 38. \$1.00.
- The sales department looks at costs.* M. J. Doohar, Editor. New York: American Management Association, 1953. Pp. 30. \$1.25.
- The American sexual tragedy.* Albert Ellis. New York: Twayne Publishers, 1954. Pp. 288. \$4.50.
- Self-perception in the university.* Edgar Z. Friedenberg and Julius A. Roth. Chicago: The University of Chicago Press, 1953. Pp. 102. \$1.75.
- The human senses.* Frank A. Geldard. New York: John Wiley & Sons, 1953. Pp. 365. \$5.00.
- Functional motor efficiency of the eyes and its relation to reading.* Luther C. Gilbert. Berkeley: University of California Press, 1953. Pp. 231. \$1.00.
- A clinical approach to children's Rorschachs.* Florence Halpern. New York: Grune & Stratton, Inc., 1953. Pp. 270. \$6.00.
- Mechanism of corticosteroid action in disease processes.* Oscar Hechter. New York: The New York Academy of Sciences, 1953. Pp. 192. \$3.50.
- Introduction to psychology.* Ernest R. Hilgard. New York: Harcourt, Brace and Company, Inc. Text Edition, \$5.75. Student Guide and Workbook, \$1.50.
- Sex ethics and the Kinsey reports.* Seward Hiltner. New York: Associated Press, 1953. Pp. 238. \$3.00.
- Religion, science and human crises.* Francis L. K. Hsu. New York: Grove Press, 1952. Pp. 142. \$3.50.
- Two essays on analytical psychology.* C. G. Jung. New York: Bollingen Foundation, Inc., 1953. Pp. 329. \$3.75.
- A speculation in reality.* Irving F. Lauks. New York: Philosophical Library, 1953. Pp. 154. \$3.75.
- Adolescence.* Marguerite Malm and Olis G. Jamison. New York: McGraw-Hill Book Company, Inc., 1953. Pp. 512. \$5.00.
- Men and unions.* John G. Mapes. New York: Group Attitudes Corporation, 500 Fifth Avenue, 1953. Pp. 36. \$1.00.
- The achievement motive.* David C. McClelland, John W. Atkinson, Russell A. Clark, and Edgar L. Lowell. New York: Appleton-Century-Crofts, Inc., 1953. Pp. 424. \$6.00.
- And lo, the star.* Margaret Aikins McGarr. New York: Pageant Press, 1953. Pp. 116. \$2.50.
- Mental health in the home.* Laurence Spurgeon McLeod. New York: Twayne Publishers, 1953. Pp. 243. \$3.50.
- Techniques of living.* William H. Mikesell. Harrisburg, Pa.: The Stackpole Company, 1953. Pp. 338. \$3.95.
- Psychoanalysis and personality.* Joseph Nuttin. New York: Sheed and Ward, Inc., 1953. \$4.00.
- Method and theory in experimental psychology.* Charles E. Osgood. New York: Oxford University Press, 1953. Pp. 976. \$10.00.
- Education and society.* A. K. C. Ottaway. New York: Grove Press, 1954. Pp. 182.



- Personality and adjustment.* William L. Patty and Louise Snyder Johnson. New York: McGraw-Hill Book Company, Inc., 1953. Pp. 403. \$4.75.
- Child psychology.* Leigh Peck. Boston: D. C. Heath and Company, 1953. Pp. 536. \$5.25.
- Conciliation in action.* Edward Peters. New London, Conn.: National Foremen's Institute, Inc., 1953. \$4.50.
- The child's conception of number.* Jean Piaget. New York: The Humanities Press, Inc., 1953. Pp. 248. \$5.00.
- Shame and guilt.* Gerhart Piers and Milton B. Singer. Springfield, Ill.: Charles C Thomas, Publisher, 1953. Pp. 86. \$3.25.
- Adrenal cortex.* Elaine P. Ralli, Editor. New York: Josiah Macy, Jr. Foundation, 1953. Pp. 165. \$4.00.
- Existential psychoanalysis.* Jean-Paul Sartre. New York: Philosophical Library, 1953. Pp. 275. \$4.75.
- The adolescent: A book of readings.* Jerome M. Seidman, Editor. New York: The Dryden Press, 1953. Pp. 798. \$4.50.
- Know your doctor.* Leo Smollar and Neil Morgan. Boston: Little, Brown & Company, 1954. Pp. 173. \$3.00.
- Father relations of war-born children.* Lois Meek Stolz. Stanford, Calif.: Stanford University Press, 1954. Pp. 365. \$4.00.
- Saving children from delinquency.* D. H. Stott. New York: Philosophical Library, 1953. Pp. 266. \$4.75.
- Handwriting: A personality projection.* Frank Victor. Springfield, Ill.: Charles C Thomas, Publisher, 1953. Pp. 168. \$1.75.
- The psychology of thinking.* W. Edgar Vinacke. New York: McGraw-Hill Book Co., Inc. Pp. 370. \$6.00.
- Cybernetics.* Heinz Von Foerster, Editor. New York: Josiah Macy, Jr. Foundation, 1953. Pp. 184. \$4.00.
- Hypnotism: An objective study in suggestibility.* André M. Weitzenhoffer. New York: John Wiley & Sons, Inc., 1953. Pp. 380. \$6.00.
- An introduction to scientific research.* E. Bright Wilson, Jr. New York: McGraw-Hill Book Company, Inc., 1953. Pp. 375. \$6.00.
- Driver characteristics and accidents.* Highway Research Board, Washington, D. C.: National Academy of Sciences—National Research Council, 1953. Pp. 54. \$90.
- Report of highway safety research correlation conferences.* Committee on Highway Safety Research. Washington, D. C.: National Academy of Sciences—National Research Council, 1952. Pp. 63.
- The field of highway safety research.* Committee on Highway Safety Research. Washington, D. C.: National Academy of Sciences—National Research Council, 1952. Pp. 42.
- I.E.S. recommended practice for residence lighting.* I.E.S. Committee on Residence Lighting. New York: Publications Office, Illuminating Engineering Society, 1860 Broadway. Pp. 44. \$1.00.
- The Social Welfare Forum.* National Conference of Social Work. New York: Columbia University Press, 1953. Pp. 365. \$5.00.
- Group report of a program of research in psychotherapy.* Psychotherapy Research Group, The Pennsylvania State College. State College, Pa.: William U. Snyder, Department of Psychology. Pp. 179. \$2.25.

## Personality Self-Assessment of Scientific and Technical Personnel

R. H. Van Zelst

*Kroh-Wagner Company*

and

W. A. Kerr

*Illinois Institute of Technology*

The reality of personality existence is multiple: there are the selves perceived by associates, and there are "selves" (complicated self) as perceived by the self. Then, too, there is the "paper-and-pencil self," the "projective self," the "under stress self," and other externally assessed selves. These selves tend to be unlike each other even for the same person because they exist within different frames of reference.

It seems a reasonable hypothesis that the individual in normal society who is best informed about an individual's personality is that same individual himself. Further, it appears plausible that many traits of his personality can be self-assessed with substantial validity (1, 4, 5).

### Method

**Rationale.** For at least three decades both psychology and psychiatry have been preoccupied with *external* assessment of personality. The validity coefficients culminating from these years of effort have been less than gratifying. In fact, for predicting such a criterion as job success they are characteristically near zero or non-existent (2, 4, 6). This accumulated experience might now well provoke the researcher to ask—"Who is likely to *know* the most about a given personality? Can a personality assess itself?" Have we overlooked the obvious?

Conventional paper-and-pencil personality tests attempt to infer a trait by measurement of operational symptoms assumed to be functions of the trait. Although no measurement

ever is direct, this approach introduces the obscuring influence of such intervening variables as shaky assumptions about symptoms and the poor reliability of symptom-type items based even on relatively sound symptom assumptions.

The directive clinical assessment approach compounds the errors of the conventional test approach by, in effect, introducing two additional intervening variables: (1) the personality of the clinician; and (2) the limited knowledge of the client possessed by the clinician.

These limitations of traditional assessment seem further to suggest rewarding validity in a self-assessment approach. The present study is based also on the assumptions that personality assessment is most valid when: (1) the emphasis is metric rather than impressionistic; and (2) trait concepts are maximized while verbalization is minimized.

**Subjects.** Subjects of this study were 514 technical and scientific personnel of the Armour Research Foundation (79%) and the Illinois Institute of Technology (21%). Their mean age was 31.9, standard deviation 9.1.

**Procedure.** A self-analysis questionnaire was constructed on the basis of Cattell's research (3) which lists definitive personality trait names based on factor and cluster analysis techniques. From this list of traits 56 trait names were selected.

Each subject was guaranteed anonymity and was asked: "Please rate yourself as compared with fellow scientists on the following traits utilizing the five point scale as follows:

as compared with other scientists, I probably am 1. much less; 2. less; 3. same; 4. more; 5. much more"—acquisitive, ambitious, etc. Each respondent also supplied age (nearest in five-year multiples), number of publications, number of inventions, and field of work. Of the subjects responding, 70% were in the field of Engineering and 30% in Physical Sciences.

*Criterion.* The criterion against which these self-assessed traits were evaluated was the summation of publications and inventions for each respondent. In other words the criterion was scientific productivity. The influence of age was held constant by means of partial correlation techniques. Mean productivity for the group was 11.4 with a standard deviation of 19.1.

## Results

Of the 667 questionnaires distributed (via campus mail with explanatory letters and return envelopes addressed to "Technical Personnel Research") a total of 514 (77%) were returned in usable form.

These 514 self-ratings on each of 56 traits were then correlated (Pearsonian) with the productivity (inventions plus publications) criterion. A second series of coefficients was then computed using the partial method on the original coefficients and holding constant the effect of age. Both series are shown in Table 1.

The original hypothesis that the self-rating approach may yield more significant validity coefficients than the traditional external-evaluation approach seems to be verified.

Table 1

Pearsonian Correlation Coefficients between Productivity and Personality Trait and between Productivity with Age Held Constant and Personality Trait \*

Trait	<i>r</i>	<i>r<sub>p</sub></i>	Trait	<i>r</i>	<i>r<sub>p</sub></i>
Acquisitive	-.23	-.30	Imaginative	.32	.41
Ambitious	.19	.25	Impulsive	.31	.39
Argumentative	.18	.23	Independent	.20	.26
Assertive	.17	.22	Inflexible	-.10	-.13
Cautious	-.15	-.20	Inhibited	-.25	-.33
Conscientious	.04	.05	Interests-wide	.16	.21
Constructive	.06	.08	Introspective	.20	.26
Contented	-.44	-.57	Leading	.26	.34
Conventional	-.41	-.53	Love-work	.25	.32
Cooperative	.20	.26	Optimistic	.20	.26
Curious	.31	.40	Original	.47	.61
Cynical	.03	.04	Patient	-.04	-.05
Easygoing	.00	.00	Painstaking	.12	.16
Eccentric	.04	.05	Persevering	-.02	-.03
Egotistical	.10	.13	Poised	.10	.13
Emotional	.02	.03	Practical	.20	.26
Enthusiastic	.30	.39	Reliable	.06	.08
Evasive	-.16	-.21	Reserved	-.08	-.10
Excitable	.15	.20	Responsible	.21	.27
Fastidious	.23	.30	Self-Confident	.27	.35
Formal	-.25	-.32	Self-Controlled	-.07	-.09
Frank	.13	.17	Sensitive	.00	.00
Friendly	.06	.08	Serious	-.12	-.16
Generous	.16	.21	Subjective	.24	.31
Grateful	.07	.09	Suggestible	.14	.18
Habit-bound	.03	.04	Tactful	.04	.05
Headstrong	.18	.23	Thoughtful	.02	.03
Hurried	.07	.09	Worrying	-.26	-.34

\* Italicized coefficients significant at 1% level.



This statement is based on the finding of 37 trait coefficients significant at the 1 per cent level, plus an additional three at the 5 per cent level. This represents 74 per cent of the coefficients at a non-chance level, a proportion rarely found in external evaluation. Sixteen (30 per cent) of the coefficients are of magnitude .30 or higher.

These more promising coefficients, ranging to .61, present an interesting self-picture of the highly productive scientists in this study. As a personality group these high producers describe themselves as being original (.61), not contented (-.57), not conventional (-.53), imaginative (.41), curious (.40), enthusiastic (.39), and impulsive (.39).

Also characteristic, but to a lesser degree, of these highly productive personnel are self-descriptions of self-confident (.35), leading (.34), not worrying (-.34), not inhibited (-.33), not formal (-.32), loves work (.32), subjective (.31), fastidious (.30), and not acquisitive (-.30).

To the extent that these scientists are representative, the more productive scientist appears to describe himself as an enthusiastic and impulsive personality which is original, and imaginatively non-conformist, not contented with reality as it is, curious as to the nature of this reality, and not fundamentally acquisitive in a selfish sense. The less productive scientists in this study possess an opposite self-description pattern.

This total pattern does not necessarily agree with the popular conception of the productive scientist. In this sample he is, for example, more subjective than objective in personality orientation; but probably this makes for the greater introspection which may be necessary for better and more original exploration and interpretation of the unknown. And our highly productive scientist is not characteristically cautious or inhibited; he is less cautious, more self-confident, and more impulsive than his less productive associates. Nor does he lurk modestly in his laboratory; he engages freely in leading behavior.

These results are consistent with some previous research on a related population (7, 8), particularly in suggesting relative selflessness of motive as a significant trait in highly productive scientists.

## Summary

A total of 514 technical and scientific personnel of the Armour Research Foundation and the Illinois Institute of Technology cooperated in an anonymous self-administered self-description report on 56 definitive personality trait names. These self-ratings were correlated with a criterion (inventions plus publications) of scientific productivity, holding constant the effect of age by partial correlation.

1. The original hypothesis that a self-rating approach to personality evaluation may yield results of greater validity than ordinarily found in the external evaluation approach is not refuted and even appears to be somewhat substantiated. Sixty-eight per cent of the validity coefficients exceed chance magnitude at the 1 per cent level.

2. As compared with the less productive, the more productive scientists in this study described themselves as more original, less contented, less conventional, more imaginative, more curious, more enthusiastic, more impulsive, and, somewhat less definitely, more self-confident, more leading, less worrying, less inhibited, less formal, more liking for work, more subjective, more fastidious, and less acquisitive.

Received July 28, 1953.

## References

1. Adams, C. R. and Lepley, W. M. *Personal audit*. Chicago, Illinois: Science Research Associates.
2. Buros, O. *Fourth mental measurements yearbook*. Highland Park, New Jersey: Gryphon Press, 1953.
3. Cattell, R. B. *Description and measurement of personality*. Yonkers, New York: World Book Co., 1946.
4. Dorcus, R. M. and Jones, M. H. *Handbook of employee selection*. New York: McGraw-Hill, 1950.
5. Pennington, L. A. and Berg, I. A. *An introduction to clinical psychology*. New York: Ronald Press Co., 1948.
6. Stagner, R. *Psychology of personality*. New York: McGraw-Hill, 2nd edition, 1948.
7. Super, D. E. *Appraising vocational fitness by means of psychological tests*. New York: Harper, 1949.
8. Van Zelst, R. H. and Kerr, W. A. Some correlates of technical and scientific productivity. *J. abnorm. soc. Psychol.*, 1951, 46, 470-475.
9. Van Zelst, R. H. and Kerr, W. A. A further note on some correlates of scientific and technical productivity. *J. abnorm. soc. Psychol.*, 1952, 47, 82.

## Personality Correlates of Social Conformity

Raymond E. Bernberg

Los Angeles State College

The author has recently introduced a scale (2) which is presumed to measure *social conformity*. *Social conformity* is defined as the tendencies of members of a society to manifest communality of attitudes and of behavior as a result of the restrictive influences of culture and society in personality development.

The scale utilizes the direction of perception technique of attitude measurement (1). It is a projective-type paper-and-pencil test. The content of the items of the scale was drawn from the following determinant areas of *social conformity*: (1.) moral values; (2.) positive goals; (3.) reality testing; (4.) ability to give affection; (5.) tension level; and (6.) impulsivity.

Examples of the type of items are:

Statistics show what percentage of men like to write things on the walls in men's rooms?

(a) 27% (b) 40% (c) 53% (d) 66%  
(e) 70%

Public Opinion Polls show what percentage of people feel it is silly to make close friendships because few people can really understand you?

(a) 30% (b) 40% (c) 50% (d) 60%  
(e) 70%

The scoring of the 37 items of the scale is based upon a weighted key determined by previous experimentation (2). Validation of the scale was determined by behavioral cri-

teria (2). The criterion groups used were adult male and female prison inmates; young male prison inmates; and regular white Protestant church-going groups; other groups used were college populations of all ages and both sexes and police officers of the Los Angeles Environs.

The scale is an attempt at approaching a dimension of personality from a different level than is usual in most personality tests. It attempts to measure an aspect of personality organization as reflected in attitudes derived from cultural and societal influences. In addition, it is an indirect method of attitude measurement.

The problem with which we are concerned is: What relationships exist between this scale and other direct measures of personality?

### Procedure

**Subjects.** The subjects utilized to relate the measure of *social conformity* (SC) to other differing personality measures are 89 female social welfare case workers and supervisors from the Los Angeles County Bureau of Public Assistance. They extend broadly as to age and work experience.

**Method.** The subjects were administered the SC scale and the Guilford-Zimmerman Temperament Survey (GZ) (4). This latter personality scale is a direct method of item questioning. The scale is broken down into ten traits which were derived by factor-analy-

Table 1  
Mean Scores and Sigmas of the Social Welfare Case Worker Group (N = 89) Compared to the GZ Women Norms and a Standard Population Group on SC

		Traits										SC
		G	R*	A	S	E*	O*	F	T	P	M	
Social Workers	M	18.2	18.4	13.4	19.9	19.2	19.7	16.8	17.9	18.6	11.8	11.1
	S.D.	5.4	4.5	4.9	5.2	6.6	5.2	5.6	4.5	5.6	4.5	8.3
Normative Groups†	M	17.0	15.8	13.7	19.6	15.5	16.8	15.7	18.1	17.6	10.8	12.5
	S.D.	5.2	4.7	5.5	6.3	5.7	5.4	4.8	4.7	4.9	4.1	6.9

\* Signif. Diff. .001 level between groups on these traits.

† N was 136 for Trait T, 300 for SC, and 389 for the remaining traits.

Table 2  
Intercorrelation Matrix of the GZ Traits and SC  
(Pearson Product-Moment Coefficients)

	G	R	A	S	E	O	F	T	P	M	SC
G	—	-.26	.57	.17	.12	-.09	-.14	.18	-.02	-.07	.10
R	-.26	—	-.03	-.21	.14	.27	.45	.27	.43	.14	-.19
A	.57	-.03	—	.54	.23	.02	-.19	.17	.13	.18	.09
S	.17	-.21	.54	—	.21	.18	.05	-.06	.23	-.06	-.21
E	.12	.14	.23	.21	—	.17	.15	-.35	.44	.29	-.19
O	-.09	.27	.02	.18	.17	—	.57	-.26	.72	.41	-.47
F	-.14	.45	-.19	.05	.15	.57	—	-.13	.50	.34	-.25
T	.18	.27	.17	-.06	-.35	-.26	-.13	—	-.03	.16	.10
P	-.02	.43	.13	.23	.44	.72	.50	-.03	—	.77	-.26
M	-.07	.14	.18	-.06	.29	.41	.34	.16	.77	—	-.15
SC	.10	-.19	.09	-.21	-.19	-.47	-.25	.10	-.26	-.15	—

sis procedures. They are: (1.) General activity (G); (2.) Restraint (R); (3.) Ascendancy (A); (4.) Sociability (S); (5.) Emotional stability (E); (6.) Objectivity (O); (7.) Friendliness (F); (8.) Thoughtfulness (T); (9.) Personal relations (P); and (10.) Masculinity (M).

A high score on any trait of the GZ indicates a "positive" quality. A high score on the SC indicates a socially undesirable degree of nonconformity.

### Results

Table 1 indicates the mean scores and sigmas of the sample obtained in this study compared to the GZ statistics for the women (4). In the case of the SC data, the sample is compared to statistics derived from a standard population (2). The social welfare case workers appear to be higher on the R, E, and O traits and similar to the normative population on all others.

Table 2 presents a matrix on the intercorrelations of the GZ traits and SC. The intercorrelations of the GZ traits derived from this study show a considerable amount of variable difference when compared to the GZ data (3) which are based upon male subjects and are tetrachoric coefficients. However, one must realize that both samples are highly select. This would indicate the possible high degree of probable variation in results one would expect of the GZ scale in terms of differential descriptive characteristics of a population.

The Gengerelli method of "factor exhaustion" (3) was used to find what factors or

subtests of the GZ scale provide maximum prediction for the SC measure but no significant increment in prediction in addition to that between O and SC was obtained.

### Summary

1. A group of 89 female social welfare case workers and supervisors were administered the *Guilford-Zimmerman Temperament Survey* and a *social conformity* scale. The purpose of the study was to find personality correlates of *social conformity*.

2. The sample obtained in this study was highly select and differed somewhat from a normative female population on the GZ scale, being significantly higher on *restraint*, *emotional stability*, and *objectivity*. The intercorrelations between the traits for this sample also differed considerably from those presented by the authors of the GZ scale.

3. The relationship between the two scales appears to be limited to the  $-.47$  correlation between the Objectivity factor on the GZ and *social conformity*.

Received July 2, 1953.

### References

1. Bernberg, R. E. The direction of perception technique of attitude measurement. *Int. J. Opin. Attit. Res.*, 1951, 5, 397-406.
2. Bernberg, R. E. A measure of social conformity. *J. Pers.*, in press.
3. Gengerelli, J. A. A method of analysis in which the factors are empirical tests. *J. Psychol.*, 1952, 33, 159-174.
4. Guilford, J. P. and Zimmerman, W. S. *The Guilford-Zimmerman temperament survey*. (Manual of instructions and interpretations) Sheridan Supply Co., Beverly Hills, Calif. 1949.



## Peer Nominations on Leadership as a Predictor of the Pass-Fail Criterion in Naval Air Training<sup>1</sup>

E. P. Hollander \*†

U. S. Naval School of Aviation Medicine, Pensacola, Florida

Studies by Williams and Leavitt (5) at the Marine Corps Officer Candidate School, by Wherry and Fryer (4) at the Signal Corps Officer Candidate School, and, more recently, by McClure, Tupes, and Dailey (3) at the Air Force Officer Candidate School have lent substantiation to the validity of peer nominations on leadership against various performance and operational criteria.

Summing up their findings, based on a factor analysis, Wherry and Fryer conclude that "Buddy ratings appear to be the purest measure of 'leadership.' . . . Nominations by class appear to be better measures of the leadership factor than any other variable" (4, p. 157). Williams and Leavitt note that ". . . sociometric group opinion was a more valid predictor both of success in Officer Candidate School and of combat performance than several objective tests" (5, p. 291). They conclude that the relative superiority of group opinion is attributable to the fact that "group members have more time to observe each other than do superior officers, they know each other in a realistic social context, and they react directly to each other's social-dominance behavior. All these are conditions favorable to informed judgment" (5, p. 291).

In addition to adequately fulfilling conditions of validity, the nominating technique has been found to meet acceptable standards of reliability. Thus, Wherry and Fryer report that ". . . the reliability of nominations after four months is outstandingly higher than that of any of the other variables upon which the test was made. This is probably

further evidence of the fact that the nominating technique has the property of early identification of the members of the group who constitute the two extremes of the leadership distribution" (4, p. 159). In a recent evaluation of peer ratings among Marine Corps trainees, an average reliability coefficient over a two-week period of .71 is reported by Anderhalter *et al.* (1, p. 26).

### Problem

The evidence supporting the validity of the peer nomination technique is clear-cut. Heretofore, however, the criteria utilized for validation have quite properly tended to be directly related to the initial character of the nomination. It has been assumed, with good reason, that peer nominations on leadership should be expected to correlate with a criterion derived from some variety of leadership behavior or performance measure. On the other hand, there exists very little research regarding the applicability of peer nominations *on leadership* to performance or operational criteria presumably unrelated to leadership behavior. It may well be that the so-called "leadership nominations" identify characteristics of the individual which relate to criteria in the spheres of cognition, or personal adjustment, or such a complex as ability to successfully solo an aircraft. With this prospect in view, the current investigation set forth to explore a fundamental relationship, that is, peer nominations on leadership, during pre-flight school, and success or failure through the whole of flight training. Fundamentally, two questions were posed: Do peer nominations on leadership during pre-flight correlate significantly with a pass-fail criterion for the entire flight training program? And, if so, how well do these nominations predict this criterion compared to other variables from the same stage of training, i.e., pre-flight?

\*Now in the Department of Psychology, Carnegie Institute of Technology.

†Grateful acknowledgment is extended to E. R. Sausser, Jr. for his valuable aid in the pursuance of this study.

<sup>1</sup>Opinions or conclusions contained in this report are those of the author. They are not to be construed as necessarily reflecting the view or the endorsement of the Navy Department.

### Procedure

A total of 268 Naval Aviation Cadets who entered pre-flight training during late 1951 were taken as a study sample. This group consisted of nine consecutively formed "sections" of about thirty cadets each. The cadets had already been preselected for the training program on criteria of physical fitness, age, minimum educational level, intelligence, mechanical aptitude, and background characteristics.

At the end of their third month of pre-flight training, each section was administered a leadership nomination form which presented the individual cadet with a list of his sectionmates from which he was asked to nominate the three men from the list *best* qualified for the hypothetical position of "student commander" and the three men *least* qualified. Furthermore, the instructions specifically stated that the nominator was to evaluate his nominees with regard to their "present and eventual success as military leaders." In this way, it was anticipated that confusion regarding the "leadership standard" to be applied would be obviated. It should be noted, too, that cadets were directed to ignore athletic ability as a factor in their nominations. This was considered to be a necessary and desirable part of the set in order to place some control on an ability which seemed likely to be closely related to physique. Nominations were weighted +3 for "highest," +2 for second highest, and +1 for third highest; similarly, weights of -3, -2 and -1 were assigned for the three corresponding "low" categories. A summation of these weights for each cadet was then taken as his leadership nomination score. The distribution of such scores yielded a unimodal and approximately symmetrical distribution. A standard score transformation was then utilized to afford a comparable index of the cadet's relative standing on leadership *within his own section*.<sup>2</sup>

In addition to this leadership score (LDR) derived from peer nominations, a number of other measures on the cadets were available from pre-flight. These were: ACE (College Level) Test scores obtained during the cadet's first week in training; Officer-Like-Qualities score (OLQ) assigned at the end of pre-flight by the officers in command to evaluate the cadet on qualities of leadership, military bearing, discipline and the like; and final pre-flight average (F.A.V.) based upon performance in all courses.<sup>3</sup>

<sup>2</sup> This technique derives substantially from one developed as part of ONR Contract No. N onr-o-3400 by Richardson, Bellows, Henry and Company.

<sup>3</sup> It should be noted that at no time were scores achieved by cadets on peer nominations available to authorities in the Training Command who assign OLQ or performance grades to cadets. Moreover, the cadets themselves did not know the ACE scores, final grades, or OLQ grades of their sectionmates at the time they made their nominations on leadership.

After a period of some eighteen months had elapsed, a follow-up of the study sample revealed that of the 268 cadets involved, 179 had passed flight training and had received their wings, 32 had failed flight training, 28 had withdrawn from training voluntarily, and the balance, 29 cadets, had been separated from the training program as a result of physical disqualification, illness, violation of contract, or some similar reason. With criterion groups thus established, a matrix of intercorrelations among the predictor variables was constructed and biserial  $r$ 's were computed for each of these variables against the pass-fail criterion.

### Findings

Table 1 presents the matrix of intercorrelations, validity coefficients, and beta weights for the four predictor variables. Among these, it is apparent that final pre-flight average (F.A.V.) predicts the pass-fail criterion at the highest relative level and with the greatest weight. This tends to reinforce the finding of an earlier study on pre-flight grades as predictors of flight performance (2). Second to final average, however, is the leadership score (LDR) which the cadet received from the nominations made by his sectionmates *before* he entered the flight phase of training, and *well over a year prior to the time he might receive his wings*. It should be noted, too, that the magnitude of the difference between the validity coefficients for F.A.V. and LDR may readily be ascribed to chance fluctuations. Superiors' ratings on qualities related to leadership (OLQ) yield a validity coefficient which is positive but non-significant statistically; its beta weight is of a relatively low order as well. Scores on the ACE Test appear to have decidedly limited predictive value against the criterion. On the whole, then, the validity coefficients and beta weights for final pre-flight average and peer nominations on leadership suggest that these two variables are of greatest relative validity among those considered from the pre-flight level of training. The multiple R obtained for these variables was calculated to be .33.

### Discussion

Considering the highly select nature of the population from which the sample was drawn, the complexity of the criterion applied, and the time differential between the predictor and cri-



Table 1

Intercorrelations, Validity Coefficients, and Beta Weights for Four Predictor Variables from Pre-Flight Against a Pass-Fail Criterion from Flight Training †

	LDR	ACE	OLQ	F.AV.	Pass-Fail Crit. ( $r_{lin.}$ )	B wt.
Peer Nominations on Leadership	—	(268) .30**	(268) .55**	(188) .50**	.27**	.207
ACE Test (College Level)		—	(239) .17**	(239) .43**	.07	-.089
Officer-Like- Qualities Grade			—	(239) .58**	.18	-.066
Final Average at Pre-Flight				—	.28**	.252
	N = 268		N = 239	N = 188	N = 211	
	*5% .12		*5% .13	*5% .15	*5% .21	R = .33
	**1% .16		**1% .17	**1% .20	**1% .27	

† In each case, the correlation coefficients reported are positive. The numbers in parentheses indicate the number of cases upon which the  $r$  is based. All validity coefficients have an  $N$  of 211.

terion variables, the multiple of .33 takes on stature. The fact, too, that under these conditions peer nominations on leadership should predict the criterion is still more surprising. While a coefficient of .27, accounting for approximately 7% of the variance, is not striking by itself, in relative terms it suggests that peer nominations at an early level of training may account for unique variance in predicting the criterion. A number of hypotheses are entertained below in an attempt to derive meaning from the obtained relationship.

In the first place, it may be asserted as a reasonable assumption that peer nominations on leadership reflect a cadet's social acceptance within the cadet group. Hence, those cadets who are low on leadership are apt to be social isolates as well. Their assimilation within cadet groups may be limited and their probability of successful completion of the total program may be diminished correspondingly. If it is further assumed, however, that such individuals are as likely to withdraw from training as they are to fail, it should follow that a validity coefficient for peer nominations taken against a *pass-withdraw* criterion should yield an approximation to the coefficient secured with the *pass-fail* criterion. A test of this hypothesis, by actual computation of this coefficient, yielded an  $r$  of .07; the hypothesis was accordingly rejected.

This leads to the consideration that perhaps a record of inadequate achievement at pre-flight, by the then *potential* failures, is of significance in determining their leadership scores. The correlation of .50 between final pre-flight average and peer nominations lends credence to the ascription of influence by the former variable on the latter. While this hypothesis is basically sound, it does not completely or satisfactorily

speak to the question posed because of the weight which peer nominations achieve independently of final average.

Another point of departure, from a somewhat different frame of reference, is that the kind of person who assimilates well in cadet groups—and who may consequently be expected to secure leadership status—may be the same kind of person who is reacted to favorably by instructors in flight training. This influence may be particularly felt when a cadet is in difficulty and is presented before a board of officers to determine whether he is to be failed from flight training. Should he impress the board favorably by certain subtle interpersonal mechanisms, he may unintentionally be accorded a more sympathetic hearing than will others. Whether before a board such as this or on the flight line, it seems probable that there is some weight introduced in favor of the more verbally fluent and socially facile individual. It is quite conceivable, therefore, that the obtained predictive quality of peer nominations might be accounted for in terms of some pervasive value through training of social characteristics such as these.

From this discussion, the points made will be seen to fall within two categories of conjecture: first, that the complex "leadership qualities," as defined by peer nominations, subsumes individual characteristics which are intrinsically related to the successful completion of flight training; second, that peer nominations tap a facet of the individual which is also perceived and reacted to by those who *evaluate* his performance in flight training. These categories certainly need not be conceived of as mutually exclusive of one another.

Whatever factors may be found to underlie the relationship between peer nominations on leader-



ship and the pass-fail criterion from flight training, it is fundamentally true that neither variable is of a simple, unidimensional structure. In order, therefore, to distill out their commonality in meaningful psychological terms, further research is indicated. It would appear reasonable to consider that the first step in such a direction should be to have nominators verbalize the criteria by which they make their judgments of leadership. Beyond this, it would also be desirable to undertake full-scale research with a peer nomination form specific to the nominator's estimate of the nominee's potential for successful completion of flight training.

In any event, it seems likely that the peer nomination technique may have utility far exceeding current practice or expectation. The "informed judgment" of group opinion might well be profitably exploited further.

### Summary

A study was conducted to determine the relationship between peer nominations on leadership during pre-flight and a pass-fail criterion from Naval Air Training. At the end of three months of pre-flight training, nine sections of Naval Aviation Cadets, a sample of 268 cases, were asked to nominate members of their section as best or least qualified for a military leadership position. Leadership scores were derived for each cadet. Three other scores were also obtained for the cadets from the pre-flight level of training: ACE Test; Officer-Like-Qualities grade (OLQ), assigned by officers in charge; and final overall pre-flight average (F.A.V.). Biserial  $r$ 's were computed for each of these variables against pass-fail criterion data from flight training. Appropriate beta weights were also derived and a multiple R calculated.

The findings of this study were these:

1. Peer nominations on leadership (LDR) predicted the pass-fail flight criterion at a significant level ( $r = .27$ ).
2. However, final pre-flight average ( $r = .28$ ) was of virtually equal value as a predictor.

3. Neither OLQ ( $r = .18$ ) or ACE Test ( $r = .07$ ) predicted the flight criterion significantly.

4. The multiple R for these four predictor variables against the criterion was .33. The beta weights obtained indicated that LDR and F.A.V. were bearing the load of prediction.

It was concluded that peer nominations on leadership, at the pre-flight level, might hold unique variance in predicting the pass-fail flight criterion. This was tentatively held to be attributable to the dual considerations that: first, peer nominations might subsume characteristics intrinsically related to success in flight training and, second, that peer nominations might tap a facet of the individual which is also perceived and reacted to by those who evaluate performance in flight training. Some implications for subsequent research were delineated with the suggestion that this technique be applied further.

Received July 24, 1953.

### References

1. Anderhalter, O. F., Wilkins, W. L., and Rigby, M. K. Peer ratings. *Technical Report No. 2*. St. Louis: St. Louis University, 30 November 1952.
2. Hollander, E. P. An investigation of the relationship between academic performance in pre-flight and success or failure in basic flight training. *Project No. NM 001 058.17.01*. Pensacola, Fla.: U. S. Naval School of Aviation Medicine, 24 November 1952.
3. McClure, G. E., Tupes, E. C., and Dailey, J. T. Research on criteria of officer effectiveness. *Res. Bull. 51-8*. San Antonio: Human Resources Research Center, Lackland Air Force Base, May 1951.
4. Wherry, R. J. and Fryer, D. H. Buddy ratings: popularity contest or leadership criterion? *Personnel Psychol.*, 1949, 2, 147-159.
5. Williams, S. B. and Leavitt, H. J. Group opinion as a predictor of military leadership. *J. consult. Psychol.*, 1947, 11, 283-291.

# The Retest Consistency of Army Alpha after Thirty Years \*

William A. Owens, Jr.

The Iowa State College

Personnel decisions are made every day which imply the long-term consistency of results obtained from group tests of intelligence. It is, however, relatively rare to be able to retest a group of adults with the identical measuring instrument originally employed and over a period of time equal to only a little less than one-half the average life span. The present paper is therefore devoted to a brief description of some results obtained under these conditions.

Table 1

Test-Retest Correlations of Army Alpha Subtests and Total Score after Thirty Years for 127 Iowa State College Freshmen Men

Subtests	$r_{t-r}$	$r_{o-e}$	$r_{11}$
1. Following directions	.30	.49	—
2. Arithmetical reasoning	.69	.77	—
3. Practical judgment	.56	.93	—
4. Opposites	.64	.93	.63
5. Disarranged sentences	.48	.87	.62
6. Number series completion	.62	.76	—
7. Analogies	.56	.96	.84
8. Information	.63	.73	.69
Total Score	.77	.97	—

The data were gathered in connection with an investigation of the effects of age upon mental abilities; this research has been reported in detail elsewhere by Owens (2). In this context, 127 males of mean age nineteen years who had originally taken Army Alpha,

\*The basic data upon which this brief note is based were gathered under the terms of a contract grant from The Office of Naval Research; however, the opinions and assertions contained herein are those of the writer and are not to be construed as official or reflecting the views of the Navy Department or the naval service at large.

Form 6, as freshmen at Iowa State College during early 1919 were retested with identical copies of this same examination during 1950.

The results of this testing and retesting have been incorporated in the two tables which follow. Table 1 shows the correlations for each subtest. Table 2 shows a condensed scatter table for Total Score. Taken together they indicate rather remarkable consistency, since the range of talent in a college population largely composed of graduates<sup>1</sup> is surely greatly restricted, and since the basic study previously mentioned suggests that the thirty-year age increment involved *did* affect individuals differentially. In each instance the retest coefficients<sup>2</sup> ( $r_{t-r}$ ) may be compared with corrected odd-even coefficients for the 1919 testing ( $r_{o-e}$ ); and, since these latter are considerably inflated by undue speeding in at least three instances, Gulliksen (1) lower limit estimates (formula 24) ( $r_{11}$ ) are also included where appropriate.

If a conclusion is warranted, it would seem to be that personnel decisions posited upon the long-term consistency of results obtained from our better intellectual tests are reasonably well founded.

Received June 26, 1953.

## References

1. Gulliksen, H. *Theory of mental tests*. New York: Wiley, 1950. Pp. 236-238.
2. Owens, W. A., Jr. Age and mental abilities—a longitudinal study. *Genet. Psychol. Monogr.* 1953, 48, 3-54.

<sup>1</sup> Mean education is 4 to 5 years beyond high school.

<sup>2</sup> All stated magnitudes are product-moment correlations computed from normalized standard scores and *not* from the grouped raw scores as plotted.

Table 2

1919 vs. 1950 Army Alpha Scores: Total Score for 127 Iowa State College Freshmen Men

1919 Score	1950 Score						Total
	50-74	75-99	100-124	125-149	150-174	175-200	
175-200							4
150-174					1	3	31
125-149				2	14	15	48
100-124		1	1	16	26	4	36
75-99		1	9	19	7		6
50-74		2	3	1			2
Total		1	1				127
	0	5	14	38	48	22	

# Reliability and Validity of the Kopas Personnel Test Battery

Philip Ash

*Inland Steel Company, East Chicago, Indiana*

The Kopas Personnel Test Battery (3) includes seven tests which purport to measure: (A) Ability to Think in Mechanical Terms; (B) Knowledge of Math and Science; (C) Preference for Non-Routine Work; (D) Ability to Get Along with People; (E) Emotional Stability; (F) Ambition; and (X) Manual Coordination. A novel characteristic of the tests (except the Manual Coordination Test) is their method of administration and scoring. The test questions are printed on cards mounted on panels. The answers to each question are printed around a dial, and the examinee indicates his choice by turning a pointer to the proper point of the dial. The tests are scored by noting the position of the pointers on the back of the board where the correct dial positions are marked. The device eliminates answer forms and "pencil and paper" administration. It also fails to provide a permanent record of item responses. Furthermore, the equipment necessitates individual administration.

Some of these tests are novel in intention and definitions, although it is not clear that the items in them (e.g., Ambition, Emotional Stability) measure what the names claim they measure.

Neither the Test Manual (3) nor a mimeographed bulletin by the author on personnel testing (4) includes any reliability data, and very sparse validity information. Only one published study, on the test's reliability (1), seems to have appeared. In it, Baxter reported that "correspondence with companies reported as users of the tests revealed little conclusive data on validity and no data on reliability."

The bulletin on the tests (4) suggests that the original validation was based on a sample of 480 employed workers, but no validity coefficients are offered.

This paper summarizes several reliability studies and one attempt to evaluate the test's validity.

## Reliability

Baxter (1) calculated both split-half and Kuder-Richardson reliabilities for a sample of 100 applicants for hourly positions at the Owens-Corning Fiberglas Corporation. Chew (2) calculated split-half coefficients for a sample of 249 male applicants of a steel mill. On a sample of steel mill applicants (Inland Steel Company) who were hired, test-retest coefficients were computed. For this sample, the retest interval was three months.

The results of these studies are summarized in Table 1. In general these results are fairly consistent, and suggest that the tests are not reliable enough for individual prediction. No reliability study is reported for the Manual Coordination test.

## Validity

The battery, including the Manual Coordination Test, was given to a sample of 88 employed plant protection officers for whom supervisory ratings were collected. For the ratings, the plant protection officers were divided into six groups: the top best sixth of the force, the second sixth, and so on down to the bottom (worst) sixth.

The score for test X, Manual Coordination, is originally a time score, ranging from about two to six minutes. Since time scores are not readily amenable to treatment as moment

Table 1  
Coefficients of Reliability for Kopas Subtests

Test	Baxter Split-Half (Corrected)	Baxter Kuder- Richardson	Chew Split-Half (Corrected)	Inland Steel Test- Retest
A	.59	.69	.66	.40
B	.77	.76	.88	.76
C	.87	.86	.93	.51
D	.59	.58	.64	.65
E	.58	.69	.72	.70
F	.88	.82	.91	.87



Table 2

Means, Standard Deviations, and Intercorrelations of the Subtests of the  
Kopas Battery and Supervisory Ratings  
(Sample: 88 Plant Protection Officers)

Test	Tests							
	A	B	C	D	E	F	X	I
A. Mechanical	—							
B. Math and Science	.60	—						
C. Ability to Get Along with People	.24	.35	—					
D. Preference for Non-Routine Work	-.07	.23	.16	—				
E. Emotional Stability	-.05	-.15	-.30	.05	—			
F. Ambition	.15	.02	.23	.00	.01	—		
X. Manual Coordination	.23	.09	.12	-.07	-.03	-.07	—	
I. Supervisory Ratings	.20	.28	.26	.25	-.15	.10	.02	—
Mean	15.1	14.8	17.5	22.5	18.1	13.8	44.6	2.6
S.D.	5.0	7.5	11.8	6.5	6.0	9.2	10.2	1.2

statistics, these scores were converted to speed scores:

$$\text{Speed Score} = \frac{1,000}{\text{time score}}$$

The intercorrelation matrix is reported in Table 2. Considered as a battery, the test intercorrelations are satisfactorily low. With the exception of the correlation of .6 between tests A (Mechanical) and B (Math and Science), these intercorrelations are generally not significantly greater than zero.

However, the criterion correlations are all equally low. A multiple correlation was calculated using the Wherry-Doolittle test selection method (5). The maximum shrunken multiple correlation coefficient obtainable, with tests B, C, and D, was .348. This is far too low for effective use as a selection device. The conclusion is warranted that the battery does not successfully predict performance rating of plant protection officers. While there is no basis for inference concerning the battery's validity for other occupations, these findings underscore the need for specific validity determinations. The unusual character of some of the tests (e.g., measuring "Ambition" as a function of the diversification of leisure-time interests) makes this even more necessary. The uncertain reliability of the tests in the battery, however, suggests that they hold little promise as effective personnel tools.

### Summary

1. In two independent samples, the internal consistency reliabilities of the six nonperformance tests in the Kopas Battery ranged from .58 to .93. In another sample, test-retest reliabilities ranged from .40 to .87.

2. For a sample of 88 plant protection officers, the seven tests in the battery were uncorrelated with one another, and correlations with Supervisory Ratings were negligible. The multiple correlation was .348.

3. These results indicate the need for careful validation, if the tests are to be used. They suggest, in view of the limitations on test reliability, that the battery may not prove to be of significant value in employee selection.

Received June 29, 1953.

### References

1. Baxter, B. Reliability and validity of the Kopas Wage Earner battery of tests. *J. appl. Psychol.*, 1947, 31, 39-43.
2. Chew, W. B. Internal consistency of items on the Kopas Personnel Tests as administered to applicants of a steel mill. Unpublished master's thesis, Purdue University, June 1951.
3. *Kopas Personnel Tests*. Cleveland: Management Personnel Corporation, 1946.
4. Kopas, J. S. The development and use of tests in modern personnel programs (mimeographed), 1942.
5. Stead, W. H., Shartle, C. L., et al. *Occupational counseling techniques*. New York: American Book Company, 1940. (Appendix V.)

## Response Reliability of the Activity Vector Analysis

James N. Mosel

*The George Washington University*

The Activity Vector Analysis (AVA)<sup>1</sup> is a new personality appraisal instrument which appears to be gaining wide popularity in industrial personnel circles. Although the test has received notice in business and industrial publications, there have been no reports in the professional psychological literature on its reliability or validity. The only study of which the writer is aware is an abstract by Dorcus and Jones (1, p. 402) of unpublished data on machinists received from the test's originator, W. V. Clarke. These data showed some validity against the criterion of supervisors' ratings.

The test consists of 81 descriptive adjectives, such as "easy-going," "high-spirited," "impulsive." The testee is requested to check all items which have ever been used by anyone in describing him, and then from the same list all items which he believes are truly descriptive of himself. He draws a line through any item which he does not understand. (For convenience of reference, the two sets of checked items will be referred to as the "other" and "self" choices.)

The scoring and interpretation of the test can be learned only by undergoing training administered by the test's originator. The results are presented in the form of a summary profile of six scores (an activity score and five vector scores), accompanied by an evaluative report covering the following topics: summary of major characteristics, environmental conditions, work requirements, social contacts, supervision required, accident potential, and a final over-all comment. This information is then related to the needs of a previously analyzed job.

In a test of this type there are two problems of reliability: the test-retest consistency of the testee ("response reliability") and the reliability of the interpreter's judgments. It is evident that without adequate response re-

liability there can be no interpreter reliability, nor can the interpretations have validity.

The present paper reports some preliminary evidence on the question of response reliability. Since the method of scoring the AVA is not available in open source, the reliability of the responses rather than the scores was the object of study.

### Procedure

Fifty-two employed adults in evening classes of a university were administered the AVA on two occasions, separated by an interval of about two weeks. Instructions at the first administration were designed to conceal the fact that the test would be given a second time.

As a measure of retest consistency, the common elements or overlap coefficient of correlation (2, pp. 120-123) was computed for each individual.<sup>2</sup> This coefficient is a measure of the extent to which an individual selects the same items on both trials; it is essentially the proportion of overlap between his first and second sets of choices. A similar technique has been used by Zubin (3) in measuring the similarity between two individuals on a check list. A value of 1 means that the same items were selected on both trials; zero means that there were no items common to both trials.

### Results

Table 1 shows the distribution of overlap coefficients for the "other" and "self" choices. There are large individual differences in consistency, the exact ranges being from .28 to .98 for the "other" choices and from .35 to .94 for the "self."

As an approximation of a reliability coefficient for the entire group, the mean overlap

<sup>2</sup> The formula is:  $r_{12} = \frac{n_{12}}{\sqrt{n_1 n_2}}$ , where  $n_1$  is the number of items chosen the first trial,  $n_2$  the number chosen the second trial, and  $n_{12}$  the number common to both trials.

<sup>1</sup> Copyright, 1950, by Walter V. Clarke.

coefficient was computed. The resulting values for the "other" and "self" choices were almost identical, .74 and .73. To the extent that these values can be considered as conventional reliability coefficients, they border on respectability; but the effect is reduced by the fact that for both sets of choices the majority of all overlap coefficients were less than .80.

It was found that there is only a moderate relationship between consistency on the "other" and "self" choices, the Pearson correlation between the two sets of overlap measures being .57.

Table 2 shows the means and variabilities of the number of items chosen on the first and second trials. It will be noticed that the mean number of items chosen increases slightly for both sets of choices on the second trial. Also, fewer items are chosen for the "self" choice than for the "other" choice; the variabilities are correspondingly smaller.

#### Summary

The AVA, an industrial personality test, requires the testee to select from 81 descriptive adjectives all those which anyone has ever used to describe him ("other" choices) and all those which he believes are truly descriptive of himself ("self" choices). The test-retest reliability of these choices was determined by means of the common elements

Table 1

Distribution of Overlap Correlation Coefficients between First and Second Trials

Intervals of $r_{12}$	Frequency	
	"Other"	"Self"
.90-.99	7	5
.80-.89	15	18
.70-.79	10	14
.60-.69	10	7
.50-.59	5	3
.40-.49	2	3
.30-.39	2	2
.20-.29	1	0
Total	52	52

Table 2

Means and Standard Deviations of the Number of AVA Items Chosen on the First and Second Trials

	First Trial	
	"Other"	"Self"
Mean	34.6	29.8
SD	17.0	13.3
	Second Trial	
	"Other"	"Self"
Mean	37.8	31.0
SD	17.8	11.6

correlation coefficient (proportion of overlap) for each of 52 individuals.

There was considerable individual variability in consistency. While the mean coefficients of overlap for the two sets of choices were .74 and .73, in both sets of choices over half of the coefficients were less than .80. Consistency on one set of choices, as indicated by the correlation between the overlap coefficients of the two sets of choices, was not closely associated with consistency on the other ( $r = .57$ ).

It would seem from these results that interpretations of the AVA might be disturbed by instability of response in an appreciable number of cases, the amount of disturbance depending on how much is staked on a given item in interpretation. There is the possibility, of course, that retest consistency itself might prove a useful indicator of personality, but this is not taken advantage of in the present method of utilization.

Received July 17, 1953.

#### References

1. Dorcus, R. and Jones, Margaret. *Handbook of employee selection*. New York: McGraw-Hill, 1950.
2. Peters, C. and Van Voorhis, W. *Statistical procedures and their mathematical bases*. New York: McGraw-Hill, 1940.
3. Zubin, J. A technique for measuring likemindedness. *J. abnorm. soc. Psychol.*, 1938, 33, 508-516.



# The Specialization Level Scale for the Strong Vocational Interest Blank<sup>1</sup>

Milton G. Holmen

*AFF Human Research Unit No. 2, Fort Ord, California<sup>2</sup>*

A report describing the development of the specialization level scales for the Strong Vocational Interest Blank and the Medical Specialists Preference Blank for use in planning a medical career was made by this writer in the recent monograph by Strong and Tucker (2, pp. 45-47). The purpose of the present paper is to describe an investigation into the possibility of using the specialization level scale for the Strong blank in the counseling of men who are not planning a medical career.

It was originally conceived that a scale might be developed for use in prediction of job satisfaction in the medical specialties. Such a scale was developed by assigning weights to those items on the Strong blank which showed differences between the responses of medical specialists as a group and the responses of physicians-in-general. Since this scale "subtracted" the interests of one group of medical men from those of another group of medical men, it was expected that medical interest would be subtracted out, leaving a scale which might measure interest in specialization in non-medical as well as in medical areas.

It might be of interest at this point to consider what kinds of items are assigned different weights in the Specialization Level scale. On quite a few items, such as the occupations of astronomer and author of a technical book, a plus-weight is assigned to liking the item and a minus-weight assigned to being indif-

ferent or disliking. On others, such as the occupations of bookkeeper, auto salesman, and bank teller, disliking the item is assigned a plus-weight, with minus-weights assigned to indifference or liking. Items in which the indifferent response is not weighted are quite common, such as the occupation of certified public accountant and the feeling toward pet canaries for which plus-weights are assigned to disliking and minus-weights to liking. Liking of social problem movies is assigned a plus-weight; disliking them is assigned a minus-weight. There are a few items, such as "chopping wood" and "pet monkeys," for which only the indifferent response is weighted. Poetry, smokers, and the study of agriculture are assigned minus-weights for liking, plus-weights for indifference, and no weight for disliking. On several items, such as the occupations of music teacher and YMCA worker and the activity of solving mechanical puzzles, plus-weights are assigned to the response of indifferent and minus-weights to disliking.

## Results

*Relationship between Specialization Level Scores and Educational Level.* Are scores on the specialization level scale related to amount of education? Such a relationship might be expected since high scores on the scale are obtained more often by persons engaged in occupations for which a considerable amount of specialized training is required. Mean scores on the scale were obtained for members of fourteen occupational groups. These means were obtained by use of the method recently reported by Strong (5) which provides mean scores on any scale for any group from the summary data on the responses of that group to each item on the Strong blank.

These fourteen occupational groups made up four subject-matter clusters. It was predicted that within each cluster the specializa-

<sup>1</sup> This research was conducted at Stanford University as a part of the Medical Specialists Research Project, under Contract No. W-49-007-MD-483 with the Surgeon General, U. S. Army. The opinions expressed in this paper are those of the writer and do not necessarily reflect those of the Department of the Army.

This study was a part of a doctoral dissertation (1). The writer wishes to express appreciation to Prof. Donald W. Taylor and Col. Anthony C. Tucker for guidance in conduct of the study, and to Dr. Edward K. Strong, Jr. for suggestions on the research and use of data from his files.

<sup>2</sup> This is a division of the Human Resources Research Office, The George Washington University.

Table 1

Relationship between Mean Specialization Level Scores and Mean Educational Levels

	Mean Speciali- zation Level Score	Mean Educa- tional Level (years)
Medical Group		
Medical specialist	50.0	20*
Physician	39.5	19*
Osteopath	34.9	17.0
Dentist	32.8	14.9
Social Science Group		
Psychologist	54.8	19.0
Social science teacher	41.8	16.4
Accounting Group		
C. P. A.	43.9	14.3
Accountant	39.6	12.3
Office worker	35.9	11.5
Physical Science Group		
Mathematician	48.8	18.8
Physicist	48.8	18.5
Chemist	45.5	16.8
Math-science teacher	42.1	16.4
Engineer	41.8	15.4

\* Educational level of group estimated.

tion level mean scores would correlate positively with the mean educational levels. The correctness of this prediction is indicated by the fact that, within each of the subject-matter clusters, the mean specialization scores were arrayed in the same order as the mean educational levels. The educational levels and specialization level scores of these groups are presented in Table 1.

The groups used for the comparisons presented in Table 1 were those used for the construction of occupational scales on the Strong blank (4, pp. 694-717). Higher education-level means would undoubtedly be obtained from present members of these occupational groups, but the general trend would probably vary little from that indicated in Table 1.

The relatively low scores of dentists and engineers may be taken to indicate that the scales measure a kind of specialization emphasizing theoretical rather than technical considerations. Study in the occupations in which the highest scores were recorded ordinarily involves more theoretical work than

those in which lower scores were obtained. Though the evidence on this point is far from conclusive, the data suggest an hypothesis for further investigation.

The mean score of psychologists is of particular interest, since it was the highest mean score obtained, even higher than that of the medical specialists. An objection may be made to considering psychologists and social science teachers as working in the same subject-matter area, but the social science teacher group was the only one at all appropriate on which data were available for making a comparison.

*Specialization Level Scores and Success in Graduate School.* The data above indicate that a positive relationship exists between specialization level scores and amount of education, but do not provide as precise an idea of what the scale measures as would be desirable. The specialization level scores of subgroups within two other occupations were therefore obtained to provide a more precise estimate of what the scale measures.

Strong blanks (1927 edition) were obtained from two groups of former students of the Stanford Graduate School of Business. These blanks had been administered to the students during their first year in the School. Seventy-five of the men who filled out these blanks were later awarded the degree of Master of Business Administration (M.B.A.). The other 75 had dropped out or failed before getting the M.B.A. The groups were matched by year to equalize any differences that may have existed from year to year in the admission policy of the School. Blanks from the classes of 1929 through 1941 were used. The mean standard scores of men not receiving the M.B.A. degree was 45.5; that of men getting the degree was 45.5.

The Strong blanks (1927 edition) of 150 chemists were also scored on this scale. The blanks used were a part of those obtained in development of the chemist scale on the Strong blank. Fifty of the blanks were from men who had Ph.D. degrees or had completed at least seven years of college training. Fifty were from chemists with Master of Science degrees, or with five or six years of college training. Finally, 50 were from chemists



with three or four years of college training. Most of this latter group held the degree of Bachelor of Science. All 150 were members of the American Chemical Society at the time of testing. None was a teacher or professor of chemistry.

The mean standard scores for the three groups of chemists on the specialization level scale were as follows: Ph.D. group, 52.2; M.S. group, 47.8; and B.S. group, 46.5. The standard deviations for the three distributions were 8.1, 8.3, and 8.1, respectively. The critical ratio of the difference between the Ph.D. group and the M.S. group was 2.65, a difference which would occur by chance less than one time in a hundred. The critical ratio of the difference between the Ph.D. group and the B.S. group was 3.20, which would occur by chance less than one time in a thousand. The difference between the M.S. group and the B.S. group was not significant (C. R. of .81), but was in the expected direction.

An investigation was made to determine whether or not the differences between groups of chemists with respect to specialization level scores might be due to differences in age between members of the groups. The mean age of the Ph.D. group was found to be 36.3 years at the time of testing; the M.S. group averaged 34.0 years; and the B.S. group averaged 35.7 years. None of the differences between pairs of these means was found to be significant.

Why should scores on the specialization level scale be related to amount of formal education for chemists but not for students of business administration? One reason for this difference may be that the two courses differ almost as much in purpose as in subject matter. Generally speaking, the more formal training a chemist receives, the more specialized that training becomes. Research for the master's thesis usually involves a minimum of six months of work on a narrow, specific phase of chemistry. At least a year is spent on a single problem in preparation of the doctoral dissertation. The purpose of the training for the M.B.A., on the other hand, is to provide a broad education for business executives, not to provide train-

ing for specialists in any one phase of business.

Research on these two groups suggests that the scale does not measure mere liking or tolerance for education, but willingness to restrict one's activities to a very narrow field. This, of course, is the very essence of specialization. One might object that training toward the doctoral degree in chemistry should not be compared with training for the master's degree in business administration. However, the M.B.A. is the highest degree ordinarily granted in the field of business administration, except to persons who plan to teach business subjects in colleges and universities. Furthermore, the difference found between Ph.D.'s and B.S.'s in chemistry was also found between M.S.'s and B.S.'s in that field, although the latter difference was less significant. Although this aspect of the research on the scale may not be considered conclusive, it appears on the basis of information available that the scale has been appropriately named.

The question naturally arises, in connection with such groups as chemists, whether the scale has any practical value. The data presented above indicate that the specialization level may be of value in predicting whether a given student who plans to enter chemistry will enjoy the narrowing and "heightening" of work required of the Ph.D. However, if this scale is to be used as a basis for the counseling of a person planning to undertake a program of graduate study in chemistry, it must provide more information on this subject than does the chemistry scale. The correlation between the chemist and specialization level scales (using the blanks of the 150 chemists discussed above) was found to be only .06, so the two scales certainly do not measure the same thing.

To test whether the specialization level scale would provide more information than would the chemist scale with respect to amount of graduate study to plan for, two comparisons of the efficiency of these two scales were made. Both comparisons involved finding the significance of the difference between proportions of overlap for the two scales (3, pp. 75-76). The proportion of



overlap indicates the proportion of persons in one group who would be classified as members of another group on the basis of scores on the scale in question. The first comparison made used the blanks of the 50 chemists described above who had Ph.D. degrees or had completed at least several years of college training. The second involved the 50 chemists who had a B.S. degree, or had completed only three or four years of college training. Both groups were used in determining the cutting points on which the proportions of overlap between them were based. For the Ph.D. group, the proportion of overlap on the chemist scale was .42 and on the specialization level scale was .34. For the B.S. group, the proportion of overlap for the chemist scale was .46 and for the specialization level scale was .36. For both of these groups, the differences between the two proportions of overlap were significant at the .01 level.

The data obtained from the blanks of chemists indicate that the specialization level scale can be used in at least one area outside the field of medicine. The data obtained from blanks of students of business administration point out the limitations in the use of the scale. It cannot be used to predict success in graduate school without consideration of what field the counselee plans to enter.

#### Theoretical Implications

The material presented suggests that a basic dimension of interests has been identified, however crudely. Within the subject-matter areas tested, the specialization level scale appears to separate those doing highly specialized work requiring long training from those doing other kinds of work. It may be interesting to compare the specialization level scale with the occupational level scale. The occupational level scale separates business and professional men from those doing other work. The specialization level scale makes a similar sort of separation within some of the business and professional groups. That it measures something different from what is measured by the occupational level scale is indicated by the fact that the correlation between the two scales, based on the blanks of

400 medical specialists, is only .07. The correlation would undoubtedly be higher if it were based on blanks of groups with a greater range of scores on these two scales, but the specialization level scale is of primary interest in groups for which the range of occupational level scores is restricted.

The material also suggests an extension of the concept of point of reference as used in the development and interpretation of vocational interest scales (4, pp. 553-576). The essential aspect of this concept as developed by Strong is that the best reference group from which to construct the scales used for scoring a given individual's blank would be one which included all, and only, the occupations the individual might consider entering. This is a somewhat idealistic definition, and practical considerations prevent use of a different set of scales for every person who takes the test. However, many of the persons taking the test are nearly enough alike to consider entering the broad group of occupations and professions represented in the  $P_1$  reference group. This reference group consists of the men engaged in the occupations college men ordinarily enter (4, pp. 712-713). The scales based on this group can be used only with respect to blanks of persons who do belong, or may be expected to belong, to the reference group on which these scales were based.

The suggestion made here is that scales can be developed which are based on differences between two levels of one group, and that scores of these scales can provide valuable information to persons not members of the groups on which the scales were constructed. Scales constructed to measure the differences between two subgroups within a single occupational or professional group may provide measures that are relatively independent of the occupations on which they were constructed.

#### Summary

The specialization level scale was developed for the Strong blank to separate medical specialists from physicians-in-general. Research reported here was undertaken to determine whether or not this scale might provide use-

ful information about other occupational groups, and thus identify specialization level as a dimension of interests comparable to occupational level.

Mean scores were obtained on the scale for ten occupational groups in three non-medical subject-matter areas and four groups within the field of medicine. Within each of these areas, the occupational groups were ranked in the same order by specialization level mean scores as by the mean educational level of their members. Further research indicated that chemists with Ph.D. degrees could be separated by this scale from those with less specialized training, but that the scale did not differentiate students who had qualified for the Master of Business Administration degree from those who had entered training for this degree but had failed or dropped out of school before receiving it.

While the evidence presented is not conclusive, it does indicate that a dimension of interests has been identified and that further

research on the nature of this dimension is merited. It indicates further that scales measuring intra-group differences may be of value for predicting with respect to occupational groups not used in the construction of the scales, provided norms are available for these other occupational groups.

Received June 30, 1953.

#### References

1. Holmen, M. G. *Vocational interest patterns of professional specialists*. Unpublished doctor's dissertation, Stanford Univer., 1952.
2. Holmen, M. G. The Specialization Level Scale. In E. K. Strong, Jr. and A. C. Tucker, The use of vocational interest scales in planning a medical career. *Psychol. Monogr.*, 1952, 66, No. 9 (Whole No. 341).
3. McNemar, Q. *Psychological statistics*. New York: Wiley, 1949.
4. Strong, E. K., Jr. *Vocational interests of men and women*. Stanford: Stanford University Press, 1943.
5. Strong, E. K., Jr. Norms for Strong's Vocational Interest Tests. *J. appl. Psychol.*, 1951, 35, 50-56.

# Interest Patterns for Certain Degree Groups on the Lee-Thorpe Occupational Interest Inventory<sup>1</sup>

Andrew H. MacPhail

*Department of Education, Brown University*

For several years men students entering Brown University have been asked to fill out the Occupational Interest Inventory (Advanced Series, Form A; Lee-Thorpe). The degree group patterns discussed here are based on scores made by 2,380 candidates for the A.B. degree, 170 for the Sc.B. in Chemistry, and 578 for the Sc.B. in Engineering. For the purposes of this study it seems reasonable to consider these three degree candidacy groups as being validation groups. Certainly this is true in the sense that students must meet specific requirements in order to be admitted to a particular degree candidacy, plus the effect of self-selection as manifested by the interest in seeking one degree rather than another.

Means and standard deviations were com-

<sup>1</sup> Published by the California Test Bureau, Los Angeles 28, California.

puted for each degree group on each part of the Inventory and the significance of the differences of mean scores made by the several degree groups on each part of the Inventory was then determined. Table 1 shows the pattern of mean scores with percentile equivalents for each of the degree groups. Some idea of the degree of overlap of groups may be inferred from the data in this table. However, of the 30 critical ratios computed 22 were found to be significant at the one per cent level.

Table 2 shows that the mean scores made by the Arts group and Engineer group differ by an amount significant at the one per cent level on every part of the Inventory. The mean scores made by the Arts group differ significantly at the one per cent level from those made by the Chemist group on seven of the ten parts of the Inventory, and on

Table 1  
Degree Group Patterns on the California Occupational Interest Inventory  
Mean Scores\*

	Arts (N = 2380)			Chemists (N = 170)			Engineers (N = 578)		
	Raw Scores			Raw Scores			Raw Scores		
	Mean	S.D.	%ile	Mean	S.D.	%ile	Mean	S.D.	%ile
Fields:									
Personal-Social	21.2	6.0	76.0	17.4	5.0	57.0	15.5	5.1	47.5
Nature	18.6	7.2	37.0	19.3	7.0	41.0	19.4	6.6	41.3
Mechanical	17.0	5.6	20.0	20.5	5.0	37.5	24.6	5.2	58.0
Business	23.5	8.3	72.5	17.6	7.0	43.0	19.0	6.8	50.0
Arts	21.6	7.3	67.0	16.1	5.9	40.5	16.7	6.0	43.5
Sciences	21.3	7.0	26.5	32.1	5.4	80.5	28.3	5.7	61.5
Types:									
Verbal	13.3	4.4	79.3	8.5	3.5	57.5	7.6	3.5	51.0
Manipulative	12.7	2.4	42.0	13.7	2.6	52.0	13.1	2.5	46.0
Computational	10.3	4.2	58.0	9.9	3.6	54.5	9.7	3.5	53.5
Level	73.4	8.8	73.0	75.0	8.3	78.3	74.5	8.4	76.7

\* Percentile equivalents are the publisher's.



Table 2

Critical Ratios (diff./PE diff.) for Arts-Chemists; Arts-Engineers; Chemists-Engineers on the California Occupational Interest Inventory

Note: All ratios are significant at the 1 per cent level, or better, except as indicated.

	Arts-Chemists	Arts-Engineers	Chemists-Engineers
Fields:			
Personal-Social	+13.8	+35.0	+ 6.5
Nature	1.9†	3.8	.26†
Mechanical	13.0	47.0	13.8
Business	+15.4	+20.0	3.4†
Arts	+17.2	+25.0	1.8†
Sciences	35.4	37.5	+12.0
Types:			
Verbal	+25.0	+48.0	+ 4.4
Manipulative	6.6	4.5	+ 4.0
Computational	+ 1.9†	+ 5.1	+ .99†
Level	3.5†	4.1	+ .96†

Level the difference is significant at the two per cent level. On Nature and Computational the differences would not be considered significantly great, according to current convention, since the level of confidence does not even reach five per cent. In terms of mean scores the Chemists are not as clearly differentiated from the Engineers as either of these groups is from the Arts group. However, one per cent confidence levels are reached on the Personal-Social, Mechanical, Sciences, Verbal, and Manipulation. On Business the three per cent level is reached but the differences

† 3.4 is significant at the 3 per cent level, and 3.5 at the 2 per cent.

‡ Not significant at the 5 per cent level.

(Note: Differences in favor of the first member of the group, such as Arts over Chemists, have a + sign in front of the critical ratio. Differences in favor of the second member of the group, such as Chemists over Arts, have no sign in front of the critical ratio.)

Table 3

Critical Ratios (diff./PE diff.) for Arts-Chemists; Arts-Engineers; Chemists-Engineers on the California Occupational Interest Inventory

C.R.*	Arts over Chemists	Arts over Engineers	Chemists over Arts	Chemists over Engineers	Engineers over Arts	Engineers over Chemists
48.						
47.		verbal			mechanical	
37.5					science	
35.4			science			
35.						
25.		per.-soc.				
20.	verbal	arts				
17.2		business				
15.4	arts					
13.8	business					mechanical
13.	per.-soc.					
12.			mechanical			
6.6			manipulative	science		
6.5				per.-soc.		
5.1						
4.5		computational			manipulative	
4.4				verbal		
4.1					level	
4.				manipulative	nature	
3.8						
3.5			level			business
3.4						
1.9			nature			arts
1.8	computational					
.99				computational		
.96				level		nature
.26						

\* All critical ratios over 3.5 in this table are significant at the 1 per cent level, or better; 3.5 is significant at the 2 per cent level, and 3.4 at the 3 per cent.

on the other four parts would not commonly be called significant.

In Table 3 the 30 critical ratios, 22 of them significant at the one per cent level, are arranged in descending order of magnitude. The specific purpose of this table is to give emphasis to the relative differential values of the ten parts of the Inventory with respect to the three degree groups, and it is a very simple matter to discover which part or parts of the Inventory have the greatest differential value and between which degree groups. Thus, for example, the table shows clearly that the three parts having the highest differential value are Verbal, Mechanical, and

Science and are effective in distinguishing the Arts and Engineer groups.

Needless to say, in practical use the mean scores for the ten parts of the Inventory (Table 1) would be rounded off to the nearest whole unit. The writer has made considerable use of this Inventory in student consultations and feels confident that the data presented here will enhance its value for such use.<sup>2</sup>

*Received December 16, 1953.*

*Early publication.*

<sup>2</sup> E. M. Hess and E. C. Allison gave valued help in the computational work involved in the conduct of this study.

## Reliability of Short Rating Scales and the Heterogeneity of the Rated Stimuli

A. W. Bendig

*University of Pittsburgh*

Two previous articles (1, 2) have reported on the relationship between the reliability of self-rating scales and the number of categories on the scale. Two types of internal consistency measures of reliability were investigated: individual rater reliability, a measure of the ability of single raters to discriminate differences between the rated stimuli, and test reliability, a measure of the individual differences between raters in consistently assigning high or low ratings to the stimuli. Since this second type of reliability is a measure of what Guilford calls the "systematic error" of raters (4, p. 273), in a rating situation test reliability becomes a measure of "rater bias," i.e., the extent of the tendency for single raters to consistently over-rate or under-rate the particular stimuli presented to them. The results of the first two studies indicated that individual rater reliability is constant for self-rating scales with 5, 7, or 9 categories, drops slightly for 11-category scales, and appeared to fluctuate for 2- and 3-category scales. Because of the inconsistent results with shorter scales, further investigation appears necessary.

In the second paper (2) one of the hypotheses suggested was that the reliability of ratings is a function of the heterogeneity of the rated stimuli. Stimuli that are distinctly different in the perceptual field of the rater should enable the rater to make simple and consistent judgments of difference between the stimuli, while stimuli that are quite similar on the rated dimension would overlap considerably and lead to disagreements between different raters as to the relative order of the stimuli on this dimension. Volkmann (7) has summarized the evidence indicating that the width of a set of rating scale categories is partially dependent on the range of the stimuli presented to the rater to be judged. The rater tends to adjust the psychological length of the scale to fit the range of the

stimuli. However, Volkmann points out (7, pp. 280-281) that this adjustment process is not completely flexible: the categories cannot be indefinitely compressed without a loss of the rater's ability to scale stimuli. This suggests that rater reliability will decrease as the homogeneity of the stimuli increases and provides the experimental hypothesis for this study.

### Procedure

*Stimuli.* In the previous study (2) 236 Ss had rated the list of 20 foods used by Wallen (8) as to preference value. From their mean ratings these foods were ranked in order from the most liked to the least liked food. Three sublists each containing 10 foods were then selected for the present study. List 1, the list containing the most heterogeneous food stimuli, was composed of the top five and bottom five foods from this ranking. List 2, of intermediate heterogeneity, was composed by selecting in a double alternation pattern 10 foods from this ranking. Foods ranked 1, 4, 5, 8, 9, 12, etc. were used for List 2. List 3, with the most homogeneous stimuli, contained the middle 10 foods: those ranked from 6 to 15. All three lists had a mean rank of 10.5 in the original ranking, but rank variances of 58.25, 34.25, and 8.25. The original list of 20 foods had, of course, a mean rank of 10.5 and a rank variance of 33.25.

*Scales.* Four lengths of scale were used: containing 2, 3, 4, or 5 categories. The three descriptive statements used in the previous study (2) were used to verbally anchor these scales. Scales with 3 or 5 categories had an anchor under each of the end categories and also under the center category. The 4-category scale also had an anchor under each end category, but the center statement was located mid-way between the two center categories. For the 2-category scale the center anchor was omitted and the two end anchoring statements used with the other three scales were placed under the two categories. The lowest category on each scale was given a numerical weight of 1, the highest category numbered 2, 3, 4, or 5, with intermediate categories numbered accordingly.

*Subjects.* The twelve combinations of three stimuli lists and four lengths of scales were



mimeographed with instructions on single sheets and randomly distributed to 278 Ss. The Ss were students enrolled in daytime sections of introductory, social, applied, and educational psychology classes at this university during the spring, 1952-53, semester. These Ss recorded their ratings on standard five-choice IBM answer sheets for convenience in the later statistical analysis. The raters were told that the researcher was investigating the adequacy of different rating scales in assessing the food preferences of college students and were requested to sign their names to the ratings.

*Analysis.* The Ss by stimuli matrix of ratings for each of the twelve sub-groups of raters was analyzed by analysis of variance procedures. From these analyses intraclass estimates of the average reliability of individual raters were obtained (3), along with the reliability with which each stimuli-scale combination (test) encouraged rater bias among the raters in each subgroup (6, pp. 93-95). Judgment of the significance of each reliability coefficient was based upon the magnitude of the *F* ratio associated with this coefficient.

### Results

The results of the twelve analyses of variance are given in Table 1 and the obtained reliability estimates are summarized in Table 2. Individual rater reliability increased approximately linearly as a function of the heterogeneity of the stimuli with the average reliabilities of Lists 1, 2, and 3 being .22, .15,

and .06. Rater reliability rose as the number of scale categories was increased from 2, through 3, to 4, and dropped slightly, but consistently, at 5 categories. To assess the significance of these findings, Kendall's non-parametric rank coefficient *W* (5) was used. Ranking the rater reliabilities across the rows in Table 2 gave a *W* of .558 which has an approximate probability of .07 of occurring by chance (5, pp. 146-147). Ranking the same reliabilities down the columns resulted in a *W* of .333 which is significant at the .01 level. Inspection of the rater reliabilities in Table 2 suggests little interaction between heterogeneity of the stimuli and length of scale, although no statistical test of such an interaction was possible. The measures of rater bias in Table 2 present a somewhat different picture. Rater bias also increased from 2 to 4 categories and slightly declined with the 5-category scale, but the results are somewhat less consistent than with rater reliability. Also, List 2 was generally the best set of stimuli in encouraging systematic rater error and List 1 the least subject to bias, but these results are manifestly a function of the length of the scale. For example, when the 4-category scale was used, List 1 was found to be most biased and List 2 least biased.

Table 1  
Reliability Coefficients and Significance Tests for Each Rating Group

List	Number of Categories	Number of Raters	Rater Reliability			Rater Bias	
			Group	Individual	F	<i>r</i>	F
1	2	24	.83	.17	5.83**	.03	1.03
	3	25	.83	.17	5.97**	.44	1.77*
	4	22	.90	.29	9.99**	.63	2.68**
	5	23	.89	.26	9.10**	.40	1.66*
2	2	20	.62	.07	2.59**	.70	3.29**
	3	25	.83	.16	5.94**	.50	2.01**
	4	21	.85	.21	6.67**	.43	1.76*
	5	21	.79	.15	4.75**	.59	2.44**
3	2	25	.50	.04	2.01*	.30	1.43
	3	27	.12	.01	1.14	.52	2.08**
	4	23	.74	.11	3.79**	.60	2.52**
	5	22	.70	.09	3.31**	.33	1.50

\* Significant at the .05 point.

\*\* Significant at the .01 point.

Table 2

Summary of Reliability Coefficients as Functions of Stimuli Heterogeneity and of the Number of Rating Scale Categories

	List	Number of Scale Categories				Mean
		2	3	4	5	
Individual	1	.17**	.17**	.29**	.26**	.22
Rater	2	.07**	.16**	.21**	.15**	.15
Reliability	3	.04*	.01	.11**	.09**	.06
Mean		.09	.11	.20	.17	.14
Individual	1	.03	.44*	.63**	.40*	.38
Rater	2	.70**	.50**	.43*	.59**	.56
Bias	3	.30	.52**	.60**	.33	.44
Mean		.34	.49	.55	.44	.44

\* Group results significant at the .05 point.

\*\* Significant at the .01 point.

Applying the same W method to the bias measures gave values of .083 (ranking across rows) and .021 (ranking down columns), neither of which is significant at the .90 level. For rater bias the interaction of stimuli heterogeneity and length of scale appears to be the most important source of variation among the reliability coefficients.

### Discussion

The results of this study have partially clarified the relation between rater reliability and scale length for short rating scales used for self-rating. Scales with 2 categories are less reliable than those with 3 categories, which confirms the results in a previous study (2). A 4-category scale yields somewhat more reliable stimuli ratings than either a 3-category or a 5-category scale, with a 5-category scale being slightly more reliable than a 3-category scale. This last statement (3 vs. 5 categories) contradicts the previous study (2), and probably only an appeal to omnipresent sampling fluctuations can reconcile this discrepancy in the results of the two studies, especially when we note that our first study (1) found no difference in rater reliability between 3- and 5-category scales. The general conclusion of no difference in rater reliability with 3- and 5-category scales appears warranted when all three studies are considered. The small, but significantly consistent superiority of 4-category scales is interesting in light of the hypothesis suggested by Jones<sup>1</sup> that rating scales with an even number of categories may yield stimulus

ratings of higher reliability. The inclusion of a center category in a scale with an odd number of possible responses may encourage the rater "error of central tendency" (4, p. 272) and reduce rater reliability. This hypothesis needs further investigation.

The hypothesis derived from Volkmann (7) that raters cannot compress their psychological reference scale to give reliable ratings of homogeneous stimuli was confirmed. The suggestion that to achieve reliable ratings the rated stimuli should cover a wide range of the rating continuum appears eminently reasonable and, fortunately, supported by the experimental findings.

The somewhat inconsistent fluctuations in rater bias noted in the Results section of this paper preclude any sweeping generalizations. Rater bias, in this investigation, did not appear to be a consistent function of either scale length or of stimuli heterogeneity.

In the Procedure section it was noted that List 1 contained stimuli drawn from the entire range of the stimuli used previously and best duplicated the stimuli variance of the original list. This may be an explanation for the slightly larger rater bias found for List 2. Thorndike (6, pp. 229-230) has pointed out that, when the responses to test items (food stimuli) are highly correlated (as they usually are with ratings), using items selected from a large range of item difficulty level (food preference) will encourage subject discrimination. Since Lists 1 and 3 contained stimuli only from the center or from the ends of the preference continuum the obtained drop in bias for these stimuli lists could be expected. However, since this explanation is *post hoc* and based upon inconsistent evidence it is somewhat unconvincing.

Two cautions must be emphasized. The results of this and the previous two studies (1, 2) can only be tentatively generalized to the rating situation where Ss are requested to report on their own feelings, preferences, prejudices, etc. Also, the results can be applied only to ratings by relatively naive raters as represented by college students. We cannot hope that our findings will be confirmed without modification when scales are used to rate more objective stimuli or are used by more experienced raters.

### Summary

Three lists of 10 food stimuli were selected so that the lists varied in the heterogeneity of the stimuli. Preference ratings were collected from 278 Ss using rating scales with 2, 3, 4, or 5 categories. Rating reliability was highest with the most heterogeneous list and with the 4-category scale and was lowest with the most homogeneous list and the 2-category scale. Rater bias results

<sup>1</sup> L. V. Jones, personal communication.

were more tentative, with the list of intermediate stimuli heterogeneity and the 4-category scale most subject to systematic rater error on the part of the Ss.

Received July 29, 1953.

#### References

1. Bendig, A. W. The reliability of self-ratings as a function of the amount of verbal anchoring and of the number of categories on the scale. *J. appl. Psychol.*, 1953, 37, 38-41.
2. Bendig, A. W. Reliability and the number of rating scale categories. *J. appl. Psychol.*, 1954, 38, 38-40.
3. Ebel, R. L. Estimation of the reliability of ratings. *Psychometrika*, 1951, 16, 407-424.
4. Guilford, J. P. *Psychometric methods*. New York: McGraw-Hill, 1936.
5. Kendall, M. G. *Rank correlation methods*. London: Griffin, 1948.
6. Thorndike, R. L. *Personnel selection*. New York: Wiley, 1949.
7. Volkmann, J. Scales of judgment and their implications for social psychology. In J. H. Rohrer and M. Sherif (Eds.), *Social psychology at the crossroads*. New York: Harper, 1951. Chapter 11.
8. Wallen, R. Food aversions of normal and neurotic males. *J. abnorm. soc. Psychol.*, 1945, 40, 77-81.



## Scholastic Achievement of Extension and Regular College Students

Alexis M. Anikeeff

Oklahoma A & M College

Should students receive regular college credit for work completed under an off-campus extension program? In order to permit a more objective answer to the question, a study was initiated to evaluate the scholastic achievement of extension students.

### Procedure

Identical personnel management examinations were administered to approximately 39 male, evening extension students and to a similar number of male, regularly enrolled students, before and after exposure to course subject matter. The same procedure was followed twice with each group. On the first occasion, the examination covered six chapters, part 5, of a standard personnel management textbook (3), and lecture material. On the second occasion, the examination again covered six chapters, part 7, and lecture material. Each examination contained 50 four-choice test items. In addition, each examination had been twice refined by rejection of items which failed to meet established internal consistency standards. About 75% of the test items on each examination covered textbook information, while 25% of the items tested knowledge of lecture material. The same individual delivered identical lectures to both groups of students, and also administered the examinations.

Although the course in personnel management is categorized as a junior level course, and the membership is predominantly composed of junior level students, approximately one-fifth of the students were seniors and one-tenth of the students were classified as sophomores. The course was a graduation requirement for five of the 39 students. On the other hand, 34 students selected the course as an elective. Not more than one-quarter of the regularly enrolled day students were veterans of World War II.

The evening extension class was composed entirely of World War II veterans. Each veteran received thirty dollars per month for

attending and being enrolled in the course. All members were enrolled under one of the following two provisions: 1. College degree program, or 2. Two-year college certificate program. The distinction between class members classified under the two programs was almost completely ambiguous. Selection of either program rested solely with the student. Extension students shifted from one program to the other program indiscriminately, and with considerable vacillation.

Both groups of students were matched according to initial performance on each of the tests. Standard errors of the means were corrected for matching on an infallible criterion in the case of a "before" and "after" comparison of the same group on the same test. All other standard errors of the means were corrected for matching on a fallible criterion; namely, initial performance on the test. Formulas for obtaining the corrected standard errors of the means were available in Guilford's publication (1, p. 196). The corrected standard errors of the means were employed in formulas used to establish the significance of the difference between arithmetic means of correlated data.

Standard errors of standard deviations were corrected for matching on an infallible criterion or on a fallible criterion as the data dictated. Peters and Van Voorhis (2, p. 143) supplied the formulas for this computation. The corrected standard errors of standard deviations were used in formulas for deriving the significance of the difference between standard deviations of correlated data.

For the purpose of securing a value to be used in the formulas for determining the significance of the difference between arithmetic means and between standard deviations of correlated data, a coefficient of correlation was obtained for each of twelve comparisons. When the number of cases in one distribution exceeded the number of cases in the distribution with which the first distribution was compared, the superfluous unmatched cases

were dropped, as suggested by Peters and Van Voorhis (2, p. 449). Approximately 12½% of the cases were dropped for one comparison, 10% of the cases were dropped for two comparisons, 2½% of the cases were dropped for four comparisons, and none were dropped for the remaining five comparisons. In all cases, the Pearsonian product-moment coefficient of correlation was employed.

Six comparisons of data were made for each of the two tests which were administered: 1. Extension students before studying vs. extension students after studying, 2. Day students before studying vs. day students after studying, 3. Extension students before studying vs. day students before studying, 4. Extension students before studying vs. day students after studying, 5. Extension students after studying vs. day students before studying, 6. Extension students after studying vs. day students after studying. For the purpose of this investigation, both presumed and actual studying are considered to be studying.

### Results

When considering the data found in Table 1, it is well to recall that the results are based upon tests which have been refined twice. In addition, it is noteworthy that each test contains 50 four-distracter questions. Under these circumstances, an arithmetic mean of 12.50 correct answers can be obtained by random guessing. A further analysis of random guessing indicates that an arithmetic mean of 18.50 must be obtained in order that random guessing can be discounted at the 5% level of confidence.

Moreover, a mean of 20.39 must be obtained to reach the 1% level of confidence.

Evidence that factors other than random guessing contribute to a naive student's test score is documented by published research and the author's analysis of test responses. As a consequence it is reasonable to assume that random guessing represents a very conservative criterion of scores which could be obtained by students who were unexposed to course subject matter. In view of this situation, the fact that the mean of extension students after studying does not differ significantly from the random guessing mean on the first experimental test, and fails to reach the 1% level on the second experimental test, is worthy of consideration by extension program administrators.

The significance of differences was derived for testing the reliability of the differences between arithmetic means, between standard deviations, and between the obtained coefficients of correlation and true coefficients of correlation assumed to be zero. A table illustrating specific details of the foregoing analysis was omitted to reduce publication costs. However, the results indicate that the differences between SD's and AM's of the extension student distribution after studying are not significantly different from those obtained from the regularly enrolled students before studying for the first experimental examination.

On the second experimental examination, the AM of extension students after studying does not differ significantly from the AM of regularly enrolled students before studying,

Table 1  
Performance of Extension and Day Students Under Varying Conditions of  
Test Administration

Sequence	Exam.	Number Cases		Arithmetic Mean		Standard Deviation		Coef. Corr.
		Day	Ext.	Day	Ext.	Day	Ext.	D & E*
Pretest	1	39	39	17.2	14.4	3.8	3.7	.97
Posttest	1	39	38	28.4	17.8	5.7	4.4	.32
Pretest	2	40	39	18.5	16.3	3.8	4.3	.98
Posttest	2	39	35	27.5	19.9	5.4	5.3	-.24

\* Day students vs. extension students.



as was true for the first examination. However, the difference between SD's is significant at the 5% level of confidence for the same comparison on the second examination. In addition, the estimated  $r$  is not significantly non-zero for the same comparison on the first examination, although the estimated  $r$  on the second experimental examination does differ from zero at the 5% level of confidence.

Both extension and regularly enrolled students obtained higher scores after studying than they did before studying, on each of the administered examinations. However, the regularly enrolled students achieved higher scores than extension students, both before studying and after studying.

### Discussion

If extension students are to receive college credit for courses offered in an off-campus program, it is reasonable to expect that the achievement of extension students should be comparable to the achievement of regularly enrolled students. Practical considerations frequently becloud the issue. Day students ostensibly exert their major effort toward the acquisition of knowledge prescribed by school administrators. Conversely, extension students in evening classes exert their major effort toward earning a livelihood. In the actual situation a considerable overlapping of goals may occur. Nevertheless, differences in goal orientation could account for differences in motivation, and in this manner be reflected in differences of achievement between extension students and regularly enrolled students.

Physiological and psychological types of fatigue probably exert their insidious influences on both the instructor and the students. The instructor frequently drives many miles, bolts down unpalatable food in unfamiliar surroundings, and talks for three hours about subject matter previously discussed during the day to heavy-lidded students who eagerly await class dismissal and reunion with their families.

Differences in educational backgrounds between extension and regularly enrolled students, as well as other factors, may also con-

tribute to the difference in achievement scores. However, despite the reasons for the differences between the achievement of extension and regularly enrolled students, if further studies support results found in this investigation, the case for granting college credit for extension work will be severely challenged. If the administrators persisted in granting college credit under these circumstances, the administrators would be honor bound to give college credit to the regularly enrolled college students solely on the basis of payment of registration fees. Under these conditions, knowledge of lecture and textbook material would be optional.

### Summary

Identical pretest and posttest examinations were twice administered to a group of extension students and to a group of regularly enrolled college students. Six comparisons of educational achievement between groups were obtained for each of two examinations.

1. The arithmetic means of regularly enrolled students on pretests did not differ significantly from the posttest arithmetic means of extension students on identical examinations.

2. The posttest mean of extension students on the first examination did not differ significantly from a mean which could be obtained by random guessing. On the second examination, the posttest mean of extension students differed at the 5% level of confidence from the mean which could be obtained by random guessing.

3. In view of the obtained results, a question was raised about the advisability of granting college credit for work performed in evening off-campus extension courses.

Received June 29, 1953.

### References

1. Guilford, J. P. *Fundamental statistics in psychology and education*. (2nd Ed.) New York: McGraw-Hill, 1950.
2. Peters, C. C. and Van Voorhis, W. R. *Statistical procedures and their mathematical bases*. New York: McGraw-Hill, 1940.
3. Scott, W. D., Clothier, R. C., and Spriegel, W. R. *Personnel management*. (4th Ed.) New York: McGraw-Hill, 1949.



## Index of Collaboration for Test Administrators

Alexis M. Anikeeff

Oklahoma A & M College

Freedom to secure appropriate information from fellow test-takers during the administration of an examination may be considered an inalienable right by some individuals. Others may denounce this procedure as a scourge upon the American educational system. Perhaps both groups will agree that from an unmoral, and a solely objective viewpoint, such a practice is undesirable because it lowers the reliability and validity of the testing process.

Proctoring examinations, distributing several forms of an examination, rearranging the same set of questions, seating test-takers at maximum distances from each other, and haranguing test-takers that virtue will triumph are methods which have been used with varying degrees of success. Concomitant with the foregoing procedures, would it be possible to develop some method or technique which could indicate the presence of collaboration on multiple-choice tests even though the act of collaboration went unnoticed by the test administrator or the proctor? The purpose of this study was to develop and test the usefulness of such a technique.

A scrutiny of examination papers submitted by two individuals who were obviously collaborating with each other during the administration of an examination suggested the feasibility of comparing the distracters selected by each individual for his incorrectly answered questions. Under somewhat similar circumstances, Bird (1) found that a comparison of incorrectly answered multiple-choice and completion questions offered definite possibilities of detecting collaboration. For the purpose of the present study, collaboration is defined as any voluntary or involuntary dissemination of information on the part of one test-taker for the purpose of improving the test score of another test-taker during the administration of an examination.

Documented knowledge indicates that random guessing would permit one question out of four questions to be answered correctly when four optional answers are presented for

each question and one of the four answers is always correct. Random guessing by definition implies a complete absence of knowledge about the subject matter tested. Therefore, if four wrong answers are substituted for four correct answers, random guessing would nevertheless permit one of the wrong answers to be selected by chance alone. However, to the extent that more than one of the arbitrarily selected wrong answers is chosen under these circumstances, something other than random guessing may be operating.

The index of collaboration assumes that within specific levels of confidence it is possible to detect collaboration between test-takers. A comparison is made of distracters selected for incorrectly answered questions by two or more individuals. Random guessing, as previously indicated, would permit one-fourth of the total number of incorrectly answered four-choice questions selected by one test-taker to be answered with identical distracters by an adjacently seated individual. A simple illustration of the foregoing situation could be portrayed by two individuals who managed to answer 20 identical questions incorrectly on a 50-question examination which employed four-choice questions and one of the optional answers was always correct. It is reasonable to believe that five of the 20 identical questions could be answered incorrectly by using identical distracters. The number of identical incorrect answers above five needed to be shared by both test-takers before collaboration is indicated can be determined by the use of a simple formula (2) for the standard error of the frequency,  $\sqrt{Npq}$ , or read from Table 1 which is based on this formula. Symbol  $N$  in the formula refers to the number of questions answered incorrectly under these circumstances.

### Procedure

An effective measure of the collaboration index's validity is an admission of guilt by individuals who have been identified as collaborators by the index. At least ten cases, involving

Table 1  
Index of Collaboration for Use with  
Four-Choice Questions

Number of Questions Wrong on Examination Paper A	Number of Identical Questions Wrong on Examination Paper B Using Same Distracters as Found on Examination Paper A Needed to Establish Existence of Collaboration at Various Levels of Confidence			
	Confidence Levels			
Key	5%	1%	.01%	.001%
4	2.7	3.2	3.8	4.4
5	3.2	3.8	4.4	5.0
6	3.6	4.2	5.0	5.6
7	4.0	4.7	5.5	6.2
8	4.4	5.2	6.0	6.8
9	4.8	5.6	6.5	7.3
10	5.2	6.0	7.0	7.8
11	5.6	6.4	7.5	8.3
12	5.9	6.9	7.9	8.8
13	6.3	7.3	8.4	9.3
14	6.7	7.7	8.8	9.8
15	7.0	8.1	9.3	10.1
16	7.4	8.5	9.7	10.7
17	7.8	8.8	10.1	11.2
18	8.1	9.2	10.5	11.6
19	8.4	9.6	11.0	12.1
20	8.8	10.0	11.4	12.5
21	9.1	10.4	11.8	13.0
22	9.5	10.7	12.2	13.4
23	9.8	11.1	12.6	13.8
24	10.2	11.5	13.0	14.2
25	10.5	11.8	13.4	14.7
26	10.8	12.2	13.8	15.1
27	11.2	12.6	14.2	15.5
28	11.5	12.9	14.5	15.9
29	11.8	13.3	14.9	16.3
30	12.2	13.6	15.3	16.7
31	12.5	14.0	15.7	17.1
32	12.8	14.3	16.1	17.5
33	13.1	14.7	16.4	17.9
34	13.4	15.0	16.8	18.3
35	13.8	15.4	17.2	18.7
36	14.1	15.7	17.6	19.1
37	14.4	16.0	17.9	19.5
38	14.7	16.4	18.3	19.9
39	15.0	16.7	18.6	20.3
40	15.4	17.1	19.0	20.6

twenty individuals, have been validated by direct admission, indirect admission; e.g., "I won't deny it, but I certainly won't admit it," or by objective observation in isolated instances when an individual was found flagrantly copying answers from his neighbor and was permitted to continue in this manner for the duration of the examina-

tion. Under these circumstances it would appear that the index of collaboration has been successful in identifying every known case of collaboration within the past two years. Unfortunately, little was known about the success of the collaboration index in the identification of the unknown cases of collaboration.

In order to further test the effectiveness of the index of collaboration, a group of 17 regularly enrolled college students were asked to collaborate with each other during a second administration of a personnel management classroom examination which contained 50 four-choice questions. As a result of lengthened summer session classroom periods, students were asked to return to the classroom 45 minutes after the beginning of the first examination for the purpose of hearing an important announcement. When the students reconvened, they were informed that they would receive an A-grade weighted equal to that of one regular examination if they would retake their previous examination under collaboration conditions. The students were told that they must collaborate with one or with several students in order to receive their reward. The students were further informed that they were in a simulated regular examination situation, and consequently, any detectable case of collaboration would be discouraged by the instructor.

Students kept a record of the number of answers which they obtained from each student with whom they collaborated. This information was retained by each student until the collaboration analysis was completed in order to insure greater objectivity of the analysis.

### Results

Detailed results are available in Table 2 where the number of identical wrong distracters are indicated as being shared by each student paired with every other student. Of the 17 students participating, collaboration could not be uncovered for seven, collaboration was found operating on the 5% level for two students, and on the .001% level for eight students.

A comparison of collaboration index analysis with the data of extent of collaboration kept by each student, Table 3, reveals that, in the experimental situation, the index of collaboration failed most significantly in the identification of student L who secured as many as twenty answers by copying from three other students. On the other hand, the index was able to identify two students, G and O, at about the .001% level of confidence when both cooperated closely with each other and secured only five answers from each other. Other cases appear to substantiate the belief

that the index is most discriminating in the identification of one-way collaboration when an individual copies a sizable number of answers from a single test paper. A substantially smaller number of answers apparently need to be shared when active two-way collaboration is involved. Moreover, the collaboration index appears unable to identify with any particularly useful degree of accuracy, the individual who copies answers from several individuals when the individuals in question fail to reciprocate his behavior. In addition, the 5% level of confidence was found much too crude for accurate identification of collaboration.

### Discussion

The collaboration index is premised upon the operation of random guessing. Consequently, the effectiveness of the index will vary directly with the degree that random guessing is operating. Since the classroom examination administered to the experimental group was refined twice, and the distracters lacking pulling-power were eliminated, it is reasonable to believe that random guessing was present to a greater extent in the experimental situation than it would have been if a non-refined examination were used.

Despite the refinement procedure, it would

nevertheless be safe to assume that all distracters do not have equal attraction values. An analysis of distracter effectiveness made by tallying the number of times a distracter is selected could suggest whether a modification or an adjustment is needed in order to secure a more accurate indication of collaboration. For example, if on a four-choice question two distracters are never chosen, and one of the remaining two answers is the correct one, then by definition an individual has only one chance of making an error. To the extent that a considerable variation is found in the number of effective distracters among the questions on the whole examination, the application of the principle of binomial expansion may prove more useful than the standardized index of collaboration.

Although the index of collaboration used in this study, Table 1, is developed for use with four-choice questions, a similar index could be developed for examinations using any other number of distracters. Moreover, in the event that a test administrator of a four-choice examination is unaware of the effectiveness of his distracters, and if an analysis of distracters is for some reason impractical at the moment, he may feel more secure in using an index based upon three-choice questions. Under these conditions he would assume that only three of the four optional answers are effectively operating in terms of attraction for any four-choice question.

The usefulness of the index of collaboration is not limited solely to identification of collabora-

Table 2  
Paired Comparison of the Number of Identical Wrong Distracters Shared by Experimental Group Members During Collaboration Examination

Student Code	No. Wrong on Exam	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
		24	22	19	17	19	19	05	22	23	05	23	26	19	15	05	22	25
A	24		10*	5	5	7	8	2	7	3	2	7	5	8	6	2	5	8
B	22			5	5	6	14†	3	15†	5	3	12†	7	6	4	3	4	4
C	19				5	3	4	1	4	2	1	5	7	4	8*	1	2	2
D	17					7	4	4	3	4	3	5	6	6	3	3	3	4
E	19						5	2	4	3	2	5	2	6	4	2	5	6
F	19							2	16†	5	2	17†	6	7	4	2	4	3
G	05								1	2	5†	2	3	3	1	5†	1	5
H	22									6	1	15†	5	6	2	1	4	7
I	23										2	7	6	3	0	2	1	3
J	05											2	3	3	1	5†	1	7
K	23												6	3	1	2	5	6
L	26													6	5	3	7	3
M	19														3	3	4	5
N	15															1	4	3
O	05																1	8
P	22																	
Q	25																	

\* Probability of collaboration 19:1.

† Probability of collaboration 9,999:1.



Table 3

Number of Answers Indicated by Students as Being Copied from Other Class Members  
During Collaboration Examination

Student Code	No. Wrong on Exam	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
		24	22	19	17	19	19	05	22	23	05	23	26	19	15	05	22	25
A	24		*					2			7			3				
B	22		§				15†		3†			†						
C	19												2		*			
D	17										8			6				
E	19						5											
F	19								†			3†						
G	05			†							1†					5†		
H	22			†			8†					†				5		
I	23															9†		
J	05							8†										
K	23						10†		†									
L	26			†										5			7	
M	19				8													
N	15		5		4	2					5							5
O	05				§													
P	22							5†			†		4	1				
Q	25								1									
						6												

Note: Student in vertical column copies from student in horizontal column.

\* Erroneously identified as collaborating at 5% level, one-way collaboration.

† Correctly identified as being involved in collaboration at .001% level.

‡ Involved in collaboration through third individual.

§ Presumed victims of 5% level erroneously identified collaborators.

tion. Allaying the suspicions of collaboration may under some circumstances prove more heartening than the establishment of the act of collaboration. For example, did the student who failed three regular quizzes earn an excellent grade on his final examination as a result of increased motivation and preparation or through collaboration? Available evidence supports the contention that the use of the index of collaboration has motivated many test-takers to become more wary of their fellow classmates during the administration of classroom examinations.

In view of the available data it is reasonable to believe that the index of collaboration is able to identify cases of large scale collaboration which could not otherwise be identified, owing to the polished skill of the collaborators. On the other hand, the index of collaboration is relatively ineffective in identifying the individual who secures information sporadically and by means of furtive glances at numerous papers which surround him during his examination period.

### Summary

An index for the identification of collaboration between test-takers during the administration of an examination was developed, and its effectiveness tested on an experimental group.

1. The index of collaboration was found reasonably effective in the identification of collaboration despite the inability of the administrator to detect its existence during the administration of the examination.

2. The index of collaboration was most effective in the identification of large scale one-way collaboration involving the copying of at least 16% of the answers from a single adjacent test-taker.

3. Two-way active collaboration was identified when only 10% of the answers were shared by two individuals.

4. Identification of collaboration was least effective when an individual copied answers from several other test-takers in a sporadic and unsystematic manner.

Received July 30, 1953.

### References

1. Bird, C. The detection of cheating in objective examinations. *Sch. & Soc.*, 1927, 25, 261-262.
2. Guilford, J. P. *Fundamental statistics in psychology and education*. (2nd Ed.) New York: McGraw-Hill, 1950.

## Group Manual Dexterity in Women \*

Andrew L. Comrey and Gerald Deskin

*The University of California at Los Angeles*

Two previous articles (1, 2) described the results of experiments with a group manual dexterity task in men. The present experiment is a duplicate of the second of the previous experiments using women college students as subjects instead of men. Although a complete description of the experimental procedures can be found in the previous reports, a brief summary will be given here.

### Procedure

Sixty pairs of volunteer women university students were given six trials on a modification of the Purdue Pegboard Assembly Task. Instead of making each successive assembly by starting with the preferred hand, the subject was required to alternate the hand used to place the first element of each successive assembly. Thus, instead of using the standard instructions for the Purdue Pegboard Assembly Task, the subjects were instructed to begin each assembly after the first one with the same hand used to place the final washer on the preceding assembly. On this individual task, each person of the pair worked on her own pegboard. The two pegboards were placed end to end, although the girls could not watch each other since a screen was placed between them.

After the six individual trials, the screen and one of the boards were removed from the table. The other pegboard was placed lengthwise between the two girls. Six more trials were taken in which the girls worked together on the assemblies instead of working individually. The first subject, for example, would place a peg in the first hole on her side of the board, after which the second subject would add a washer and the first subject would follow with a collar, and finally the second subject would complete the assembly with another washer. Instead of repeating this operation, however, the subject who finished the first assembly would begin the next assembly by placing a peg in the

second hole on her side of the board. The first subject would then place on the first washer, and so on. Thus, the assemblies formed a zigzag pattern down the board, first one subject beginning the assembly and then the other, functions alternating each time. In terms of the kind of set required, this process appears to be similar to the individual assembly task in which the subjects were required to alternate hands in starting successive assemblies.

On the basis of the sum of scores for the last four trials on the individual task, the members of pairs were divided into two categories, "high" and "low." Reliabilities for "high," "low," and "group" scores were determined by correlating scores on trials three and five with scores on trials four and six and correcting by the Spearman-Brown formula. "Difference" scores were also obtained by subtracting the "low" total score of each pair from the "high" total score. All intercorrelations were computed between "high," "low," "difference," and "group" scores. The correlations of the "high" and "low" scores with the "group" scores were corrected for attenuation in both variables and beta weights computed for predicting "group" scores from "high," "low," and "difference" scores.<sup>1</sup> The multiple correlation coefficient was also computed.

<sup>1</sup> In the two previous experiments, the "difference" scores were not included in the analysis. It was decided to add them as a factor of possible influence and to give a check on the whole process of obtaining beta weights. Since the "difference" scores are completely determined by the "high" and "low" scores, the whole system is overdetermined for computing beta weights so that only an approximation to a solution is usually obtained. The errors in reproducing original correlations from the partial regression equations will be small if the correlations are internally consistent. These errors are recorded under the "e" column in Table 1. The "high-low" correlation was not corrected for attenuation here, as in the previous analysis, because this correlation should not increase as the proportion of error variance decreases. Data from the previous experiments have been reworked so that Table 1 gives the proper figures for all three studies based on the method of analysis decided upon for the present experiment.

\* This research was supported by a grant from the University of California.

In the two previous experiments, as well as this one, the main objective has been to determine the extent to which performance of persons on a group task could be predicted from a knowledge of how well they could do individually on a very similar kind of task. Corrections for attenuation were used to gain some knowledge of the theoretical upper limit of this predictability.

### Results

The results of this experiment and the two previous experiments with men are summarized in Table 1. The results for the men are given in rows marked "I" and "II," respectively, whereas the figures for the women in the present experiment are given in the rows marked "III." In the first column of Table 1 are listed the total score categories, "high," "low," "difference," and "group"; means and standard deviations for these sets of scores are given in the second and third

columns, respectively. The reliabilities are given in column four, with the intercorrelations of score variables appearing in the next four columns. Beta weights are given in the next column, and in the last column are given the differences between experimental correlations with "group" scores and those reproduced by inserting beta weights and correlations in the partial regression equations (see footnote 1).

The pattern of results for university women is much the same as that for university men except that the over-all multiple correlation is only .62 as compared with .66 and .70 in the first two experiments, respectively. The differences are not significant, however.

The most important fact which emerges from these three experiments is that a surprisingly small proportion of the total variance on a group-performance task can be predicted on the basis of how well the team members can perform individually on an ap-

Table 1

#### Summary of Three Experiments

Note: Experiments I, II, and III were with 65 pairs of male graduate and undergraduate psychology students, 47 pairs of male undergraduate psychology students, and 60 pairs of female undergraduate psychology students, respectively. The numbers under the "e" column are the discrepancies between obtained correlations with group scores and those correlations which result from placing the best beta weights in the partial regression equations. In the three experiments, the multiple correlations were .66, .70, and .62, respectively, with the squares equal to .44, .49, and .38. The  $r_{11}$  values for the difference scores were computed as the geometric mean of the "high" and "low" reliability estimates.

Score		Mean	S.D.	$r_{11}$	Correlation with				Beta Wt.	e
					High	Low	Diff.	Group		
High -	I	192	16.5	.90	1.00	.48	.48	.50*	.35	.01
	II	156	18.0	.89	1.00	.46	.51	.47*	.35	.02
	III	160	13.3	.91	1.00	.54	.59	.42*	.29	.03
Low -	I	173	16.8	.92	.48	1.00	-.56	.53*	.41	.01
	II	137	17.2	.94	.46	1.00	-.49	.53*	.47	.01
	III	142	13.0	.79	.54	1.00	-.42	.48*	.41	.02
Diff. -	I	18.9	17.4	.91	.48	-.56	1.00	-.06	.00	.00
	II	21.2	19.1	.92	.51	-.49	1.00	-.03	.00	.02
	III	18.4	12.9	.85	.59	-.42	1.00	.00	.00	.02
Group -	I	178	19.2	.87	.50*	.53*	-.06	1.00		
	II	186	19.2	.77	.47*	.53*	-.03	1.00		
	III	190	18.8	.86	.42*	.48*	.00	1.00		

\* These correlation coefficients were corrected for attenuation in both variables before they were used to obtain the beta weights.



parently similar kind of task. In none of the three experiments did that proportion reach one half. These figures are based on correlations corrected for attenuation, and, as such, represent what could be done with errorless measures. Practically, the possibilities for prediction are even less impressive.

These results suggest that there are other important behavior variables to be measured which will help to determine how well a person will perform in cooperative kinds of tasks. Evidently a careful analysis of the physical operations he performs in the group task, followed by measurement of these job elements

by means of individual tests, leaves out important sources of variance. Just what the nature of these other sources of variance may be is not immediately evident. Further research will be needed to explore this problem.

*Received August 3, 1953.*

### References

1. Comrey, Andrew L. Group performance in a manual dexterity task. *J. appl. Psychol.*, 1953, 37, 207-210.
2. Comrey, Andrew L. and Gerald Deskin. Further results on group manual dexterity in men. *J. appl. Psychol.*, 1954, 38, 116-118.

## A Short Method of Factor Analysis<sup>1</sup>

Robert C. Wilson

*Reed College, Portland, Oregon*

and

Andrew L. Comrey

*The University of California at Los Angeles*

This article will point out the need for a short method of factor analysis in certain kinds of situations, describe briefly such a method, and finally present an empirical comparison of the short method with the complete centroid method on a particular example.

Occasions frequently arise in psychological research where several variables have been developed over a period of time to assess certain characteristics of individuals, groups, and situations. The variables may be scores on individual items, groups of homogeneous items, or other types of data. At some point, revision may be needed to yield a smaller number of variables which will cover the domain with equal effectiveness and greater economy. This economy may be achieved through the elimination of variables which are measuring only those functions already assessed elsewhere, and through the retention of relatively uncorrelated measures.

Where any considerable number of variables is involved, factor analysis constitutes a useful technique for providing the information upon which such a program may be undertaken. Unfortunately, the most generally used factor-analysis procedures are so time consuming that they are often by-passed in situations where they could be helpful. To factor analyze the items of a test, for example, using the complete centroid method becomes quite an expensive and time-con-

suming assignment if the test contains as many as thirty or forty items.<sup>2</sup> Or, if we are dealing with about the same number of questionnaire variables, each variable being assessed by means of a pool of homogeneous items, a similar situation prevails. Even though a factor analysis might be valuable, then, it may not seem feasible to take on the amount of labor required in the methods traditionally employed.

Thurstone (1) has described a diagonal method of analysis which is considerably less laborious than a complete centroid solution, especially when the problem of rotation is taken into account. Thurstone feels that the diagonal method, while theoretically correct, is inadequate as a method of analysis because, generally, the communalities which are inserted in the diagonals cannot be very accurately estimated. Since the mechanics of the method place a rather heavy dependence upon the diagonal cell values, inaccuracies in these values may produce considerable distortion in the final result.

Although it is difficult to obtain accurate estimates of the communalities, it is frequently possible to compute good reliability estimates for the tests. Since test reliabilities can be estimated more accurately than test communalities, the accuracy of a diagonal method of analysis can be improved by substituting reliabilities for communalities in the diagonals. This changes the nature of the problem to some extent, however. Instead of analyzing only the common factor variance, or some estimate of the extent of that vari-

<sup>1</sup> This research was carried out under Contract N6-ONR-23815 between the University of Southern California and the Office of Naval Research. The research was supervised by R. C. Wilson. The modified diagonal method of factor analysis herein discussed was conceived by A. L. Comrey. The opinions expressed are our own and are not necessarily shared by the Office of Naval Research. J. M. Pffinger, J. P. Guilford, and H. J. Locke are the co-responsible investigators.

<sup>2</sup> An iterative method has been suggested (3) for factor analyzing test items. The method described in this paper could be employed effectively in the preliminary stages of such an analysis for the purpose of shortening the iterative process.

ance, the total true variance is used, including what would ordinarily be assigned to specific factors if communality estimates were used.

Many factor analysts would object to analyzing the total test space in that they are looking for factors representing underlying variables which "explain" the *common* elements among the variables concerned. If this is the objective, of course, it would be unsatisfactory to use the diagonal method modified by using reliabilities in the diagonals instead of guessed communalities. On the other hand, careful analysis of the purposes in factoring may reveal that the objective is not at all that of discovering some latent explanatory structure. There may be, for example, several well-developed scales which need to be refined and extended, but which cannot be discarded just because a factor analysis solution suggests that the "best" variables are somewhere in between the ones actually in use.<sup>3</sup>

The problem under such circumstances is more nearly one of imposing a relatively arbitrary structure upon the domain rather than attempting to develop a new set of variables of special intrinsic explanatory value. When this is the case, there is no good reason why the first factor should not be aligned with the best developed and most useful variable. The next factor can go through or near a variable of established value which is approximately orthogonal to the first one.

In this way, the factors can be made to coincide in so far as possible with variables which are already serving adequately at the time. Other variables fall where they may, and are eliminated as they fail to add anything to the structure needed to lay out the domain under investigation. As a result of such an analysis, some variables may need to be refined in certain directions to make them more independent of one another. Other variables can safely be dropped altogether in

that they are not adding anything new and the factorial picture will very likely suggest new areas in which further variables may profitably be developed.

The considerations just presented led the authors to attempt an empirical check of the diagonal method using reliabilities in order to gain some information concerning how it might compare with the complete centroid method. The results of that empirical check will be presented after a brief description of the mechanics of the method.

### The Method

Thurstone (1) has described the diagonal method fully, but a brief repetition of the essential steps may be helpful:

1. Compute the correlation matrix as usual, inserting test *reliabilities* in the diagonal cells. This is at variance with the diagonal method described by Thurstone (1) in that guessed communalities would conventionally be used.

2. Take the square root of the reliability which is largest and divide every correlation in the corresponding column by this number. The resulting quotients are factor loadings for Factor I.

3. The variance due to Factor I is removed from the correlation matrix by obtaining the matrix of inner products of the first factor loadings and subtracting them from the original correlation matrix, including the diagonal values. Thus, if tests 2 and 3 had loadings of .6 and .7 in Factor I,  $.6 \times .7 = .42$  would be subtracted from the original correlation between tests 2 and 3. The result of this operation would be entered in the matrix of first factor residuals.

4. The entire process is repeated, each time taking the column with the highest remaining diagonal value, until the unextracted variance is presumed to be largely error. This will usually be evident when the square roots of diagonal entries begin to get small in relation to residual column entries, resulting in obviously inflated factor loadings. The exact point at which factor extraction should cease, however, must remain to a considerable extent a matter of judgment.

5. The matrix of factor loadings obtained

<sup>3</sup> Thurstone (2) reports a study of Guilford's temperament schedules, in which he wished to know how many factors were represented in the 13 scores, rather than in ascertaining their common factors. For this purpose he used reliability estimates in the diagonal cells and factored the matrix by the centroid method.



in this fashion is ready for such rotations as may be necessary to align the factors satisfactorily. Since the factors have been extracted in a manner designed to make them similar in nature to certain existing variables, the number of rotations required to finish the solution generally will be rather small.

### Results of Two Analyses

A previous article (4) reported the results of a complete centroid analysis of certain questionnaire variables used by the authors in the study of supervision and employee attitudes in a naval shipyard. A modified diagonal method of analysis was also applied to the same correlation matrix, using reliabilities instead of guessed communalities, to determine the extent of the discrepancies which might occur between the results of the two different procedures. The previous article treated the nature and significance of the centroid factorial results, so that we will be concerned here only with comparing the two solutions, rather than in presenting the findings of the factorial process.

In the centroid analysis, nine factors were extracted. Seven factors were interpreted, leaving two residual factors. Of the seven interpreted factors, three were factors specific to doublet variables included in the analysis. In the modified diagonal analysis, seven factors were extracted, two of which were doublets. The two solutions achieved excellent agreement on six factors, leaving each with a doublet factor not clearly present in the other analysis. Table 1 presents the comparative findings for the two studies. Loadings are listed only where the value in one or the other analysis was .40 or more. For information on the nature of the variables involved, the reader is referred to the previous report (4).

### Discussion

The total amount of labor expended in carrying out the modified diagonal analysis was approximately one-eighth of that required to complete the full centroid analysis, including the rotations. Much of the saving is in the rotation process itself, since only a

minimum of readjustment of the axes is necessary in the modified diagonal case. In the present example, 14 graphical orthogonal rotations were required for the modified diagonal analysis, and 79 such rotations were carried out with the centroid factors.

The agreement between the two analyses, as evidenced by the data in Table 1, would suggest that the loss in accuracy is not great in comparison with the time saved, provided the over-all objectives of the analysis are consistent with a sacrifice of this kind.

The 25 variables for this analysis were derived from 13 dimensions or item pools by dividing each dimension into comparable halves or sub-dimensions on the basis of item content. One sub-dimension was discarded because of lack of item homogeneity. The correlations between comparable halves were used as the reliability estimates to be inserted in the diagonal cells. For 15 of the variables this correlation between halves was their highest correlation. Since the highest correlation in the column was used as the communality estimate in the centroid analysis, 15 of the diagonal cell values were the same for both analyses. This favored a similar outcome in both analyses.

Inspection of Table 1 reveals that many of the differences in loadings between the two analyses may be attributed to discrepancies in the amount of variance extracted. These differences in extracted variance are revealed by comparison of the sums of squares of factor loadings for each variable in the two analyses. Variables 1 and 2, for example, had low diagonal entries for the modified diagonal analysis because the reliability estimate obtained by correlating variables 1 and 2, which were supposedly comparable half-dimensions, was low. Evidently the low reliability estimate was due, in some degree, to lack of comparability between halves; the communalities for the centroid analysis were higher. Had higher values been inserted instead of the reliability estimates actually used, the loadings for variables 1 and 2 would have been higher. It is expected that agreement of factor loadings using these two methods will be greater for variables where the

Table 1  
Comparative Factor Loadings\*

	I <sub>c</sub>	I <sub>d</sub>	II <sub>c</sub>	II <sub>d</sub>	III <sub>c</sub>	III <sub>d</sub>	IV <sub>c</sub>	IV <sub>d</sub>	V <sub>c</sub>	V <sub>d</sub>	VI <sub>c</sub>	VI <sub>d</sub>	VII <sub>c</sub>	VII <sub>d</sub>	h <sub>c</sub> <sup>2</sup>	h <sub>d</sub> <sup>2</sup>
1															51	31
2													40	13	45	32
3			78	85									42	27	76	77
4			83	84											74	76
5			49	50			40	53							60	66
6			38	44	25	44	51	55							67	73
7							51	57	31	40					62	64
8							50	66							60	74
9					51	87									58	1.05
10					55	58									46	43
11			44	46											57	46
12			44	43											59	42
13	70	74			45	35									84	84
14	65	70			45	45									81	86
15	87	77					47	21							1.18	90
16	65	75													72	79
17	65	65													68	60
18	65	61													64	74
19	68	61											24	46	77	79
20	55	63											25	44	73	68
21	73	68													78	75
22	21	50														
23									70	57					82	70
24									56	60					55	53
25											68	69			64	52
											65	75			57	70

\* Questionnaire variables are numbered down the left side of the table. Factors are numbered across the top, with solutions of the centroid (c) and diagonal (d) analyses placed side by side for each factor. Factor loadings are shown only in those cases where the loading obtained by one of the two methods is .40 or more. Decimal points have been omitted.

amount of common factor variance approaches that of the true variance.

### Summary

Occasions arise where it is desirable to apply factor analytic techniques, but the exploratory nature of the work and the time available may not justify a complete centroid analysis. A diagonal method, modified by using reliabilities instead of guessed communalities in the diagonal cells, is suggested as a satisfactory and economical substitute for the complete centroid analysis under certain conditions. The results of an empirical comparison of this method with the complete centroid method on one correlation matrix

show that the two agreed fairly closely upon most of the factors obtained.

Received July 21, 1953.

### References

1. Thurstone, L. L. *Multiple factor analysis*. Chicago: The University of Chicago Press, 1947.
2. Thurstone, L. L. The dimensions of temperament. *Psychometrika*, 1951, 16, 11-20.
3. Wherry, R. J., Campbell, J. T., and Perloff, R. An empirical verification of the Wherry-Gaylord iterative factor analysis procedure. *Psychometrika*, 1951, 16, 67-74.
4. Wilson, R. C., High, W. S., Beem, H. P., and Comrey, A. L. A factor-analytic study of supervisory and group behavior. *J. appl. Psychol.*, 1954, 38, 89-92.

## A Methodological Study of Cigarette Brand Discrimination

Richard A. Littman

*University of Oregon*

and

Horace M. Manning<sup>1</sup>

*University of Minnesota*

It is a common belief that many consumer's goods are identical, at least in the sense that they are indistinguishable once the wraps are removed. Recent studies by Pronko and his colleagues (3, 4) seem to have demonstrated this with respect to cola beverages; the case for cigarettes seems to have been settled the other way in one set of investigations (1, 6) and positively in a recent study (5). However, many of the studies are not conclusive because of errors of procedure or analysis.

The reasons for objection differ for the different studies. They reduce, however, to a question of the type of judgment asked of Ss. For example, let us consider the studies of Pronko (3, 4). In these studies, various cola brands were administered to Ss who were asked to identify them by name. In two studies using popular varieties of cola, discrimination was random, i.e., names were applied on a chance basis. Consequently, in a third study three obscure brands were used. Not once were they properly identified; instead the names of the major brands were applied, also in a random fashion. The authors conclude that "the seven brands of Cola beverages employed in our series of studies appear to have the same stimulus function for our subjects and may be said to be 'equivalent stimuli'" (3, p. 608).

This conclusion seems unwarranted, for even if there were a discriminable difference among the different colas, Ss could hardly express their awareness of this difference if they were unfamiliar with the names to be applied to them. It may be that the colas do indeed have equivalent stimulus values, but the use of an identification, as in this case, does not settle the case for this conclu-

sion. If Ss apply names with better than chance accuracy, one may conclude that discriminable differences exist; but random results do not justify the conclusion that such differences do not exist. For example, what would the results have been if Ss were asked to respond by "same" or "different"? There is no *a priori* reason to expect comparative judgments to yield the same results as identification judgments.

This importance of the type of judgment used suggests that discriminatory ability in this sort of study is a function of test procedures as well as test materials. The field of psychophysics provides ample evidence that this is the case. The questions raised here merely spell out the well-founded generalizations concerning the relationships between discrimination and procedures.

In the present study, two specific questions were raised: (1) Is there evidence that individuals can discriminate among different cigarette brands? (2) Are there differences in the patterns of judgment for two different kinds of discriminations, *viz.* recognition and affective?

The idea of comparing affective and recognition judgments is based on the following consideration. While the ability to apply a name correctly to something requires some specific training, the ability to say whether one likes or dislikes something ordinarily does not. To be sure, likes and dislikes may be radically changed as a result of experience, as may the willingness to say one likes something. Even the very nature of the qualitative experience may be altered as a result of such experiences. It seems, nevertheless, that at any given time almost any object can be responded to affectively, even in the absence of past contact with that object.

It may be this very universality that per-

<sup>1</sup> The data were collected by Mr. Manning for a Master's thesis at the University of Oregon.



mits affective experiences to play such a great role in behavior; on one's first contact with some particular object, say a wine or tobacco, the only reaction he may have available is an affective one. With this in mind, it was hypothesized that *one* of the things that determined preferred brands would be one's affective response to them, and therefore one's usual brand would tend to be liked more than other brands. For any given sample of objects, this does not exclude the possibility that other objects may be liked even more, or that all of these objects may be equally liked.

While this hypothesis might well have been wrong if any differential association between brands and affective judgment appeared, the use of affective judgments would still merit consideration as the possibility of a new way of approach to some common problems of discrimination. Indeed, the "method of impression" lies at the heart of psychophysics. However, the use to which it is put here is somewhat unusual since it is used as a "sensitivity" test rather than a simple preference test. Between the limiting cases of 100% "like" and 100% "dislike," judgments of this sort might be used where it is difficult or impossible to develop in Ss any other system of reporting their discriminations.

Originally we had conceived of comparing the recognition and affective judgments for accuracy. This has proved possible only in the roughest way because of the logical problems attending the definition of "accuracy" for the affective judgments. In a sense, affective judgments cannot be veridical instruments; one may have an affective experience, but one does not have a *correct* affective experience. But one can properly seek to correlate such experiences with known differences like preferences or brands. So, when we speak of accuracy the reader may substitute "sensitivity" in the case of affective judgments without serious harm to the thread of the discussion. If this phase of the study is thought of as a methodological inquiry into different response procedures permitted Ss, then while accuracy may be asserted only of the recognition judgments, sensitivity may be asserted of the results of such judgments in terms of differential association with some

known differences, *viz.* brand preference and brand differences. In other words the *methods* are being studied in terms of their sensitivity or accuracy, not the judgments themselves.

Technical problems prevented the study's being carried out with colas so that one cannot say anything about cola discrimination on the basis of these results. But the logic of the argument should be generalizable to discrimination studies in any modality, and that is our main objective.

### Procedure

*Materials.* Three brands of cigarettes were used: Camel, Chesterfield, and Lucky Strike. These were selected because a preliminary survey indicated that they accounted for about 78 per cent of the brands used by a sample of 239 students at the University of Oregon. Philip Morris ran a close fourth, but was not used because of the difficulty in concealing identifying print. It will be seen that the method of investigation is not seriously affected by this omission. The students involved in the survey were not the same ones who participated in the main study.

Each cigarette was banded with a gummed label applied in a tacky state. While this method, which was also used by Ramond, Rachal, and Marks (5), undoubtedly removes some cues usually used by smokers, e.g., "So round, so firm, so fully packed . . .," it was felt that the procedure of blindfolding presented more serious difficulties. Each S, therefore, smoked a cigarette whose upper half was covered by a paper label, but whose lower half was completely normal.

*Subjects.* There were 288 Ss culled from introductory psychology classes. There were somewhat more male than female Ss, though the data have not been analyzed by sex. They were recruited by having the following instructions read at the beginning of the class hour:

"We are conducting an investigation in cigarette judgments, and would like your cooperation. Will all of you who are regular cigarette smokers raise your hands? . . . Will you file out individually by rows, so that when one returns, the next may leave, etc. The experiment, which is not unpleasant, should take about five minutes of your time. Thank you very much."

For reasons made clear under *Routine*, 246 of these Ss were utilized in the analysis.

*Routine.* Ss were divided into two groups by alternating assignments as they entered the experimental room. Each S was greeted in the following manner:

"Come in. Just have a seat here, please. Now, I'm going to give you a cigarette and I'd

like you to put it in your mouth, and when you're ready, I'll light it for you. (*E* hands cigarette to *S*.) Take three or four puffs—more if you like—just as you normally would.

(*A*) then tell me whether you like or dislike it; or

(*B*) then tell me whether you think it is or is not the brand you usually smoke.

"Ready? . . . Here's your light." (*E* lights cigarette for *S*, waits till *S* indicates judgment is made, then takes cigarette and records response.)

"Do you have a package of cigarettes with you? May I see it? Is this the brand you usually smoke?" (*E* records response.) "That's all and thank you very much. Since the success of this experiment depends in part upon people not knowing what to expect when they enter, I'd appreciate it if you didn't discuss the procedure with your friends."

Those *Ss* assigned to the *affective* group were given the (*A*) portion of the instructions; the *recognition* group was read the (*B*) statement. In all other respects treatment was identical.

There are four things especially to note about the instructions and the procedure.

1. *E* did not know what brand of cigarette *S* smoked in the experiment until the latter had left the experimental room. The cigarettes had previously been masked and placed together. *E* arbitrarily selected one from a paper sack and gave it to *S*.

2. Each *S* made only one judgment. Previous studies have erred seriously in this respect. The statistics available for analyzing situations of this sort invariably call for independence of measurements. The repetition of tests upon the same *S* is frequently unanalyzable because of the impossibility of computing any meaningful coefficient to express the relationship between their judgments. In addition, it seems unlikely that the usual methods of adapting *S* by having him wash out his mouth between drinks or tasting a mint between puffs are entirely satisfactory, unless one can be sure that there is no cumulative effect of the neutralizer. Finally, if the materials to be discriminated are in fact discriminable, it must be demonstrated in some independent way that the "trace-reducer" does not have a differential effect upon the trace. If such a differential interaction actually exists, none of the previously reported studies is designed so that its effect could be properly evaluated as a source of error.

3. *Ss* indicated their reaction by stating only whether they liked or disliked the cigarette or whether it was the brand they usually smoked. The reason for this technique is as follows. A like-dislike judgment is disjunctive; recognition judgment in the form of a naming response would not be, though it could be converted into one. In order to keep the response as uniform as possible, however, *Ss* were asked to make a disjunctive identification. It should be noted that they were not given any information con-

cerning the brands offered to them. For all they knew, they might have been getting "off brands." The similarity of this portion of the procedure to that of Pronko and his colleagues is obvious. Not giving *S* a set of possibilities to draw on prevents the computation of a chance level of success; this will be more fully discussed in the *Analysis*.

4. After *S* made his judgment he was requested to exhibit a pack of cigarettes, and indicate whether it was the brand he usually smoked. As a result of this step, 42 *Ss* were eliminated from the analysis; either they did not smoke one of the three brands selected for study, or they had no cigarettes with them. While some regular smokers were undoubtedly excluded by this tactic, the likelihood of getting the occasional smoker seems small. As a matter of experimental procedure, it seems fair only to include individuals who may be expected to show the maximal degree of discrimination, i.e., the regular smoker or drinker. Finally, the use of this device ensured complete obscurity of the test stimuli from the *Ss*; to have indicated preference for smokers of certain brands alone would have cancelled the precautions to make the recognition judgment as comparable to the affective one as possible.

*Test Locus.* For all but 15 *Ss*, the experiment was carried out in the same room. It was high-ceilinged, with a large window, a single ceiling fixture, two chairs and two tables. The room was thoroughly aired out before the next subject was admitted. All work was done during daylight hours. The remaining 15 *Ss* were studied in another room similarly furnished. Since they were distributed evenly (save for one) between the two groups, if there were any measurable effect attributable to the difference between the rooms it could serve only to increase the variance of the two distributions. It was decided, therefore, to leave this possible source of variance in the error term of the statistic.

## Results

In Table 1 the results for the recognition and like-dislike judgments are presented. In each cell, the top number represents the total number of "yes" or "like" responses, respectively, and the lower number, the total number of observations in this category. Thus the bottom numbers represent the total of "yes" plus "no" responses, and of "like" plus "dislike" responses, respectively, in each category. As will be seen, the consigning of any observation to a particular category depends on three variables, *viz.* the brand of cigarette *S* regularly smoked, the brand he sampled, and which type of judgment he was asked to make.

Table 1  
Frequency of Judgments

			Recognition		
			Brand Preference		
			Ca	Ch	LS
Brand Sampled	Ca	Yes	7	3	3
		Total	17	11	15
	Ch	Yes	4	3	2
		Total	16	10	16
	LS	Yes	3	3	7
		Total	16	9	16
			Like-Dislike		
			Brand Preference		
			Ca	Ch	LS
Brand Sampled	Ca	Like	10	4	4
		Total	14	9	16
	Ch	Like	7	5	5
		Total	17	10	14
	LS	Like	8	3	10
		Total	14	9	17

It is readily seen that the design of this study permits an analysis of variance technique. Therefore a three-way analysis was carried out. Independence of observations is assured by having only one judgment per S. The raw data are frequencies and are unusable in that form. They were converted into a relatively normalized distribution by means of the arcsine transformation (2, 6). While there are variations in the number of observations per cell, comparison of the theoretical vs. the obtained residual variances indicated little damage to the resulting analysis.

Table 2 shows the data converted to proportions, and Table 3 summarizes the resulting analysis of variance. A problem always arises concerning the proper error term in evaluating the significance of the various effects. We have tested for *AB*, *AC*, and *BC* interaction by the *F*-ratio, with *ABC* interaction as the denominator. *AC* and *BC* are not significant at the five per cent level, while *AB* seems high enough to warrant another test. Consequently we have pooled the *AC*

and *BC* interactions with the *ABC* interaction in order to obtain more degrees of freedom and a more powerful test. The *F* test for *AB* interaction then becomes:

$$\frac{136}{25.75} = 5.28, \text{ against } F_{0.05}(4, 8 \text{ df}) = 3.84,$$

while for *C*-effect we have:

$$\frac{571}{25.75} = 22.17, \text{ against } F_{0.05}(1, 8 \text{ df}) = 5.32.$$

Since the *AB* interaction is significant, the appropriate test for the *A*-effect is:

$$\frac{56.5}{136} = 0.415, \text{ against } F_{0.05}(2, 4 \text{ df}) = 6.94,$$

while that for the *B*-effect is:

$$\frac{42.5}{136} = 0.312, \text{ against } F_{0.05}(2, 4 \text{ df}) = 6.94.$$

The following significant effects emerge: *AB* interaction and *C*. What does this mean? The interaction of brand preference and brand sampled, *AB*, results in a change not attributable to either variable alone. That is, *Ss* *did* like their own brands more often or identify them as their own more often than they did other brands. One may conclude from this that there exist differences in brands such that regular users of these brands can distinguish among them.

Table 2  
Proportions of Judgments

		Recognition		
		Brand Preference		
		Ca	Ch	LS
Brand Sampled	Ca	0.412	0.273	0.200
	Ch	0.250	0.300	0.125
	LS	0.187	0.333	0.437
		Like-Dislike		
		Brand Preference		
		Ca	Ch	LS
Brand Sampled	Ca	0.714	0.445	0.250
	Ch	0.412	0.500	0.357
	LS	0.571	0.333	0.588



Table 3  
Analysis of Variance—Three Way

	Sum of Squares	df	Mean Square	F-ratio
A-effect (Brand Sampled)	113	2	56.5	
B-effect (Brand Preferred)	85	2	42.5	
C-effect (Recognition vs. Like-Dislike)	571	1	571.0	
AB	545	4	136.0	4.39*
AC	78	2	39.0	1.26†
BC	4	2	2.0	0.0065†
ABC	124	4	31.0	
Total	1,520	17	89.4	

\* Vs.  $F_{0.05}(4, 4) = 6.39$ .

† Vs.  $F_{0.05}(2, 4) = 6.94$ .

But what of the other hypothesis outlined, viz. that the affective judgment technique is more sensitive or accurate than the use of recognition judgments? In certain respects the findings are surprising. The proportions for the two kinds of judgments are different, i.e., the ratios of "like-dislike" and "my brand: yes-no" responses are different. These different ratios cannot be directly compared, however, to determine the relative accuracy of the two kinds of judgments. In order to do so it would be necessary to determine some level of chance expectancy; this cannot be done for the present design where Ss were unaware of the possibilities available to them.

However, it is possible to re-analyze the data in order to estimate the *relative* accuracy of the two kinds of judgments. This is done by constructing a two-way table for each of the judgments separately. The results are shown in Table 4. It will be seen that A- and B-effects are not significant in either of the analyses. This is similar to the finding in the three-way table above, and again indicates no significant effect due either to brand sampled or brand preference alone. Now, the three-way analysis indicated that the AB interaction was present. There is no way of evaluating the AB interaction for the two-way analyses, but they can be evalu-

ated against one another:

$$\frac{AB \text{ (like-dislike)}}{AB \text{ (recognition)}} = \frac{97}{70.2} = 1.38,$$

against  $F_{0.05}(4, 4 \text{ df}) = 6.39$ .

We conclude that there is no difference between AB interaction on like-dislike or recognition judgments, i.e., neither one is significantly superior as a means of distinguishing between brands. It should be pointed out that there is a tendency for the like-dislike judgments to be superior. With samples as large as these, however, it hardly seems large enough to be of practical significance.

### Discussion

In the introduction, the use of a naming procedure was taken to task. The essence of the argument, it will be recalled, was that even if two or more stimuli were appreciably different to S, he might not be able to express his awareness of this difference if the task demanded of him be one of applying names to the stimuli. To test this, Ss were presented with either a recognition or an affective assignment. Both seemed about equally accurate.

Now it is certainly true that to name something is to identify it. However, to identify something does not require that it be named. In other words, the process of recognition is broader than that of appellation. In either case, it is clear that there is some other process which we may call discrimination upon which recognition rests, logically if not temporally. The suc-

Table 4  
Analysis of Variance—Two Way

	Sum of Squares	df	Mean Square	F-ratio
Recognition Judgments				
A-effect	21	2	10.5	0.149*
B-effect	61	2	30.5	0.434*
AB interaction	281	4	70.2	
Total	363	8	45.4	
Like-Dislike Judgments				
A-effect	170	2	85.0	0.876†
B-effect	28	2	14.0	0.144†
AB interaction	388	4	97.0	
Total	586	8	73.6	

\* Tested against AB interaction for recognition.

† Tested against AB interaction for like-dislike.

cuss of the affective judgment attests to this. There is, to be sure, the likelihood that distinctions between test objects which are based upon affective judgments involve entirely different kinds of discriminations from those based upon recognition. The fact that the proportions of the two kinds of judgments differed indicates that we did not have a single process. The tendency for like-dislike to be superior is further support along these lines, though the difference is obviously of theoretical rather than practical import, being of such a small magnitude. It is further likely that different kinds of recognition, e.g., naming or "sorting," involve different processes in part. After all, to name something properly requires a greater amount of precision in using the various cues provided by the stimulating object or event. But it is obvious that a sound investigation in an area of this sort requires that such distinctions be kept in mind.

The study seems to have demonstrated, also, that one need not be restricted to "cognitive" reactions to objects in order to test for discrimination of an object's properties. To be sure, affective reactions tap different properties; but in the field of discrimination where in field situations the exact basis for the discrimination is often unknown, their use seems to be another possibility. Identification, recognition, etc. may have an affective base as well as the usually assumed sensory base. It then becomes a fascinating problem to tease out the role of the two factors in any given series of discriminations and most challenging to determine what cues are responsible for the affective judgment. In any case, if a substantial correlation between the two types of judgments can be demonstrated (as seems most likely), one may have a rapid technique for determining an S's sensitivity to differences where this approach is applicable.

There remain, as always, many questions concerning the type of judgmental situation that should be used. The answer to such questions, we believe, can best be formulated after the objective of the investigator has been stated as precisely as possible. At such a time the particular variables to be manipulated should emerge more clearly. In any case, it is obviously desirable that a person be *capable* of making a naming reaction before such a naming or other differential reaction is used to decide whether a discrimination is possible. Similar errors can be avoided if one thinks of discrimination as being partly a function of materials, partly of procedure, and if one thinks of it, further, as existing in various degrees of precision or exactness.

### Summary

It was proposed that the use of a recognition judgment, as employed in previous studies of cigarette and cola brand discrimi-

nation, is not the most appropriate test of discriminability. Therefore, in the present study the use of a recognition judgment and an affective (like-dislike) judgment were compared. A total of 246 regular cigarette smokers were divided by alternation into two groups. Members of one group made a recognition judgment, the other a like-dislike judgment. Each S was given one of the "Big Three" cigarettes with brand name obscured.

Results, analysed by analysis of variance using the arcsine transformation, were as follows:

1. Both types of judgment were made with better than chance accuracy.
2. The like-dislike judgment technique was slightly more sensitive than the recognition judgment, but not significantly so.
3. The distribution of responses (dichotomous in both cases) for each type of judgment was radically different, suggesting that, while about equally sensitive as applied to the present problem, each is an expression of a different type of psychological function.
4. It was suggested that the use of an affective judgment may have a greater applicability to problems of discrimination than it has enjoyed, and merits further study in this respect.

Received July 7, 1953.

### References

1. Husband, R. W. and Godfrey, J. An experimental study of cigarette identification. *J. appl. Psychol.*, 1934, 18, 220-223.
2. Mood, A. F. *Introduction to the theory of statistics*. New York: McGraw-Hill, 1950. P. 346.
3. Pronko, N. H. and Bowles, J. W., Jr. Identification of cola beverages. (In three parts.) *J. appl. Psychol.* (1) 1948, 32, 304-312; (2) 1948, 32, 559-562; and (3) 1949, 33, 605-608.
4. Pronko, N. H. and Herman, D. T. Identification of cola beverages. IV: Postscript. *J. appl. Psychol.*, 1950, 34, 68-69.
5. Prothro, E. T. Identification of American, British, and Lebanese cigarettes. *J. appl. Psychol.*, 1953, 37, 54-56.
6. Ramond, C. K., Rachal, L. H., and Marks, M. R. Brand discrimination among cigarette smokers. *J. appl. Psychol.*, 1950, 34, 282-284.
7. Snedecor, G. W. *Statistical methods*. Ames, Iowa: Iowa State University Press, 1940. Pp. 380-383.

# Pointing Accuracy of a Joy Stick Without Visual Feedback

W. D. Garvey and W. B. Knowles<sup>1</sup>

Naval Research Laboratory, Washington, D. C.

The following problem was suggested by a practical situation in which it was necessary to know how accurately an operator could point a joy stick at a target without visual reference to the position of the stick. The problem as it was investigated may be defined as determining man's ability to point a joy stick at a series of small points of light displayed in the space of an otherwise totally dark room.

## Procedure

**Apparatus.** A general pictorial representation of the apparatus is presented in Figure 1. The joy stick was an aluminum rod, one-half inch in diameter, 6.5 inches long, mounted on the shafts of two potentiometers. When the stick was pointed at a target, the horizontal and vertical components of the stick's position were converted by the potentiometers into voltage readings, which were calibrated in terms of degrees of deviation in azimuth and elevation.

The joy stick was located on a pedestal in the center of the dark room 68 inches from the forward wall, above the floor, and below the ceiling. The Ss sat in a chair on a platform immediately behind the pedestal so that the joy stick was at about stomach level; they were instructed to grasp the stick in the right hand, palm up and thumb extended forward along the top of the stick.

The targets were 24 stationary small lights located at predetermined positions about the room. Since the room was totally dark and the brightness of the targets was very low, Ss were able to detect the target with little knowledge of the relative distance involved. The targets may be regarded as lying on the surface of a sphere of unspecified diameter, the center of which was located at the joy stick. When the joy stick was pointed directly ahead of S it indicated the zero-zero point for the azimuth and elevation dimensions of the targets. This center point (C in Figure 1) was denoted by a red cross (illuminated between trials only) which was mounted at the center of the forward wall, 68 inches from the floor. The positions of the targets were located in terms of degrees of azimuth and elevation from point C. These positions are given in Table 1. The schema of these positions may be interpreted with the aid of Figure 1; e.g., target No. 9 (labeled as such in Figure 1) was located 16° below and 16° to the right of point C. The position of each of the other targets may be similarly interpreted.

**Method.** The Ss were seven Naval enlisted men stationed at the Laboratory to serve as subjects; all Ss were right-handed. Without previous dark adaptations Ss would have had better visual acuity at the end of an experimental trial than at the beginning. Since it was desired to maintain constant the visual component of the

*Method.* The Ss were seven Naval enlisted men stationed at the Laboratory to serve as subjects; all Ss were right-handed. Without previous dark adaptations Ss would have had better visual acuity at the end of an experimental trial than at the beginning. Since it was desired to maintain constant the visual component of the

Table 1  
Target Positions and Response Errors  
N = 28

Target No.	Target Position* (Degrees)		Response Errors (Degrees)	
	Elevation	Azimuth	Mean	S.D.
1	-74	+56	10.3	7.2
2	-74	-45	10.4	6.7
3	-46	+79	16.0	9.5
4	-46	+16	13.9	7.1
5	-46	-11	15.0	7.8
6	-46	-74	15.7	7.6
7	-14	+76	12.8	5.8
8	-15	+44	11.8	7.3
9	-16	+16	8.8	4.6
10	-16	-14	9.9	6.0
11	-14	-45	8.7	4.9
12	-14	-74	10.7	7.3
13	+16	+75	16.9	10.3
14	+15	+45	11.9	5.7
15	+15	+15	12.3	8.2
16	+15	-15	10.9	7.5
17	+15	-42	12.4	6.0
18	+15	-73	13.9	7.3
19	+47	+75	20.5	10.0
20	+47	+15	13.6	8.0
21	+47	-15	11.7	6.6
22	+47	-74	17.9	11.3
23	+75	+45	18.2	9.4
24	+73	-40	17.1	7.8

\* A plus sign indicates upward elevation and right azimuth; a minus sign indicates downward elevation and left azimuth.

<sup>1</sup> The authors wish to acknowledge the valuable assistance of Mr. Manus Munger and Mr. Gerald Hasson, formerly at this Laboratory, who assisted with the experiment and carried out the major portion of data analysis.



task, Ss were given 20 minutes of dark adaptation before an experimental session.

The task was considered to be a relatively unfamiliar one for Ss. Therefore, before the experiment proper, each of the seven Ss was given two practice periods (one per day). This practice amounted to four trials of pointing the stick at each of the targets.

The experiment proper began on the day following practice. The Ss were given four experimental trials, two per day for two successive days. An experimental trial consisted of one presentation of each of the 24 targets. The targets were presented to Ss in a different randomized order each trial.

The S was given as much time as was needed to make the pointing response, and was instructed to report when he considered the stick to be pointing at the target. The E immediately recorded the position of the stick, extinguished the target light, illuminated the center cross, and then instructed S to return the stick to the center position. The next aiming response was initiated by having S move the stick from this centered position to a position of pointing at a new target. Thus each aiming response began

and ended by having the stick pointed at the center cross.

The S's pointing responses were measured in terms of degrees of elevation and azimuth deviation of the stick's pointing position from the target's position. This elevation and azimuth deviation was later transformed into a great circle deviation, which was a measure of the direct angular displacement between the pointing position of the stick and the position of the target.

The Ss were never given any knowledge of the correctness of their responses or the direction of their errors.

## Results

*Magnitude of Errors.* The mean error for all Ss and all targets was  $13.4^\circ$ , with a range for single stimuli from  $1^\circ$  to  $52^\circ$ . Table 1 presents the magnitude of errors to specific targets and the respective standard deviations. Mean errors with respect to location are presented in Figure 2. The data indicate that response errors to some targets were greater than to others. The smallest errors appear

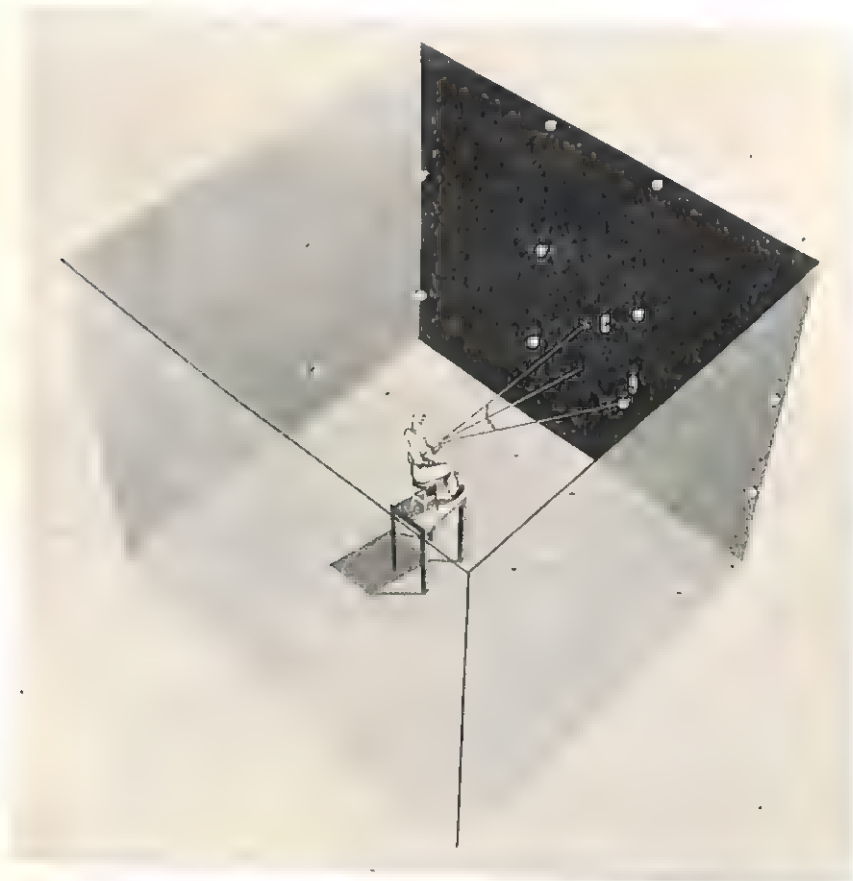


FIG. 1. Schematic representation of experimental situation. C = center point; 9 = position of target No. 9; S and hand stick are located in center of room.

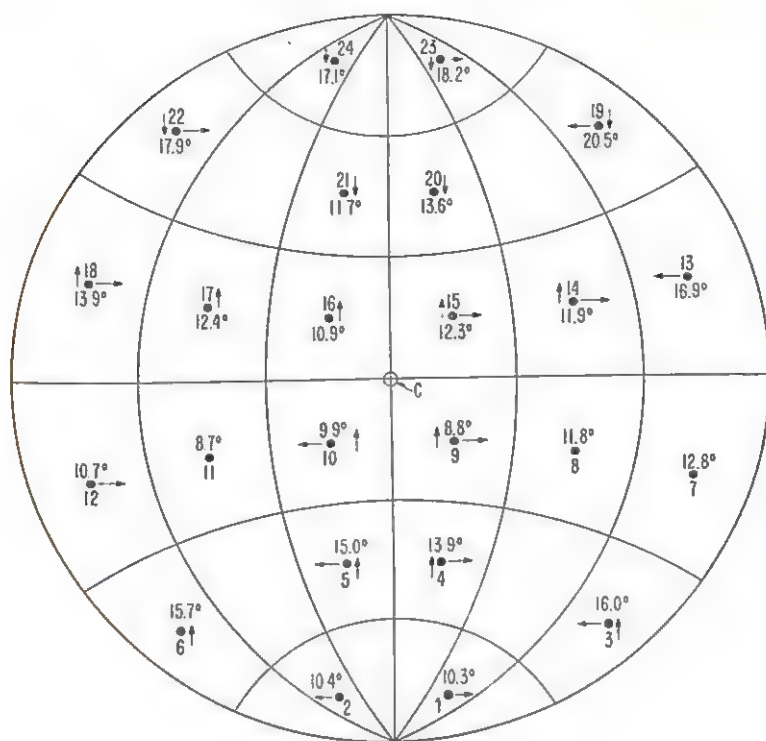


FIG. 2. Diagrammatic representation of magnitude and direction of response errors. Mean errors, in degrees of arc deviation from target, are presented for each target relative to center point (C); significant directional error deviation is indicated by direction of pointed arrow.

to have been made to targets just below the center of the room (i.e., targets 7 through 12, all approximately  $15^\circ$  below the center) and to the extremely low targets (i.e., targets 1 and 2, approximately  $75^\circ$  below the center). Generally speaking, the largest errors were made to targets located on the extreme right and above the center of the room; next largest errors were made to targets on the extreme left and above. Statistical analysis<sup>2</sup> indicated that response errors to targets above the center of the room were significantly greater ( $p < .05$ ) than those below the center; in addition, there was a propensity for right response errors to be greater than left response errors ( $p < .10$ ).

**Direction of Errors.** There was a tendency for Ss to err in a particular direction for specific targets. Statistical sign tests<sup>3</sup> were made on the direction of Ss' response errors.

The arrows in Figure 2 give an indication of the direction of the errors made to a specific target; for example, the upward pointing arrow at target No. 4 indicates that a statistically significant ( $p < .05$ ) number of Ss' responses were in the direction of the stick pointing above the target. The right-left arrows may be similarly interpreted. There were two directional tendencies in elevation errors. For responses made to targets below the center and  $15^\circ$  above the center there was a tendency for Ss to aim above the targets. However, for those targets  $45^\circ$  or more above the center there was a tendency for Ss to aim below the targets. There also appear to be two directional tendencies in azimuth errors. In general, Ss tend to aim to the left of targets located on the extreme-right and to the right of targets on the extreme left; i.e., Ss are disposed to undershoot the targets on the extremes. However, there is also a tendency for Ss to aim to the left of targets located just to the left of the center of the room and

<sup>2</sup> Dixon, W. J. and Mood, A. M. The statistical sign test. *J. Amer. statist. Ass.*, 1946, 41, 557-566.

<sup>3</sup> See footnote 2.

to aim to the right of targets located just to the right of the center of the room. For these targets located around the center, Ss appear to overshoot the targets.

### Discussion

The results indicate that more than 90% of Ss' pointing responses were within  $25^\circ$  of the targets. In most applied situations, where joy stick control is feasible, such accuracy is adequate. It is clear, however, that pointing accuracy is partially determined by the locus of the targets. The fact that Ss were required to hold the stick in a particular fashion is certainly an influential factor in determining the differential pointing accuracy for the various targets. Pointing at targets located on the extreme right and upward required a more "strained" muscular response on the part of Ss with the result that responses to targets located in these areas would be more difficult responses to make. From interviews with Ss after the experiment it was learned that even though S could not see the joy stick or his arm, he attempted to point the stick at the targets as if he were sighting down the right arm; i.e., S was pointing the entire arm as well as the stick at the targets. Such aiming was not possible with all targets, but it was Ss' belief that when such aiming was possible, they were able to respond with more ease and accuracy. Such a mechanism operating in the pointing procedure would have facilitated responses made to targets to the left and downward. The data imply that these two factors, manner of hand grip and aiming as if with the entire right arm, may

have influenced pointing accuracy, for responses towards targets which were downward and to the left were more accurate than those upward and to the right.

Although statistical analysis indicates that no improvement in performance took place during the course of the experimental trials, the fact that Ss respond with consistent error biases would indicate that Ss may learn to point more accurately if they are given knowledge of their results during the course of practice.

### Summary

An experiment was conducted to determine how accurately man can point a joy stick at visual targets without visual feedback as to the position of the stick. The results of this experiment may be summarized as follows:

1. Pointing errors ranged from  $1^\circ$  to  $52^\circ$  with a mean of  $13.4^\circ$ . Ninety per cent of the pointing errors were  $25^\circ$  or less in magnitude.
2. There was a correspondence between magnitude of errors and locus of the visual target. Errors were largest for responses made to targets located to the right and above the center of the room. There was a tendency for responses to be more accurate if they were made to targets either to the left or below the center of the room.
3. There was a tendency for Ss to undershoot targets located around the periphery of the target space and to overshoot targets located around the center of the space.

*Received July 16, 1953.*



## Rate Accuracy in Handwheel Cranking \*

Robert S. Lincoln

*The Johns Hopkins University*

When a rate of movement is produced by a human operator in an attempt to reduce error in tracking performance, accuracy of control may be limited by the inability of the operator to maintain a steady speed. His rate of movement may not continuously match the required rate. Because of this limitation it becomes important to determine the accuracy with which various rates of movement can be maintained.

In a handwheel cranking task it is possible to introduce variation in the required rate of movement in three ways. 1. The required angular speed of rotation may be increased or decreased for a given radius of movement. When this is done, the required linear rate also increases or decreases. Linear and angular rates, therefore, change in combination. Linear rate in this case refers to the units of distance traveled per unit of time by the handwheel knob. 2. The required linear rate may be changed while holding the angular rate constant. This is accomplished by varying the radius of movement for any given angular rate. 3. The required angular rate may be changed while the linear rate is held at a constant value. This is accomplished by making compensating adjustments in the radius of movement as angular speed changes.

The effects of variation in both linear and angular rates of movement have been studied with regard to performance in manual tracking tasks. Helson (3) has measured the accuracy obtained with various rates of movement and different radii of cranking in compensatory tracking. Lincoln and Smith (4) have varied the angular and linear rates in combination in a direct-pursuit tracking task. The relationship between rate of movement and accuracy is complicated, however, by the nature of the tracking devices.

The accuracy of direct-pursuit tracking depends upon the combined accuracy of both rate and positioning movements (4). This results from the fact that accurate tracking is achieved only when the operator matches the position of the target with a specific position of the handwheel knob and simultaneously matches the rate of target movement with a proportional rate of handwheel motion.

Consistent relationships between the position of the target and the position of the handwheel knob are eliminated with compensatory tracking devices, but in both compensatory and pursuit tasks it is possible to achieve the required rate of movement without maintaining the alignment of cursor and target. For this reason tracking error records do not give an accurate picture of the operator's ability to maintain a steady rate of speed. Furthermore, in both of the studies described, changes in the required angular rate of cranking were produced by changes in gear ratios. This procedure reduces the load on the handwheel as rate increases and may affect the relationship between speed and accuracy.

### Purpose of this Experiment

This report is concerned with the accuracy with which different linear and angular rates of cranking movements can be maintained for clockwise and counterclockwise directions of turning.

In order to eliminate the difficulties inherent in tracking devices and to study rates of movement in greater isolation, a special task has been devised. Four characteristics of the task are of particular importance. 1. The subject is presented with a display consisting of a target and cursor. The cursor instantaneously indicates the rate of handwheel movement by its spatial position relative to the target. It is impossible for the subject to achieve the required rate of movement without aligning the target and cursor. Errors are recorded as errors in rate. 2. With the

\* This work was supported by Contract N5-ori-166, Task Order I, between the Office of Naval Research and The Johns Hopkins University. This is Report No. 166-I-178, Project Designation No. NR 145-089, under that contract. Miss Frances Wolf-ram aided in the collection and analysis of the data.

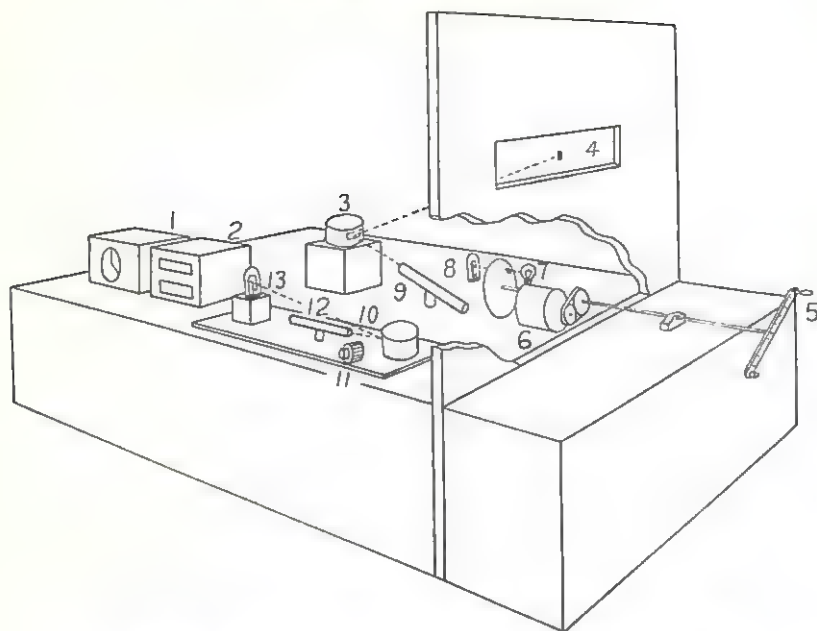


FIG. 1. Schematic diagram of the apparatus: (1) accuracy-recording clock; (2) electric counters; (3) display galvanometer; (4) screen; (5) crank; (6) tachometer generator; (7) light source for revolutions counter; (8) photoelectric cell; (9) light source; (10) recording galvanometer; (11) variable resistor; (12) light source; (13) photoelectric cell.

apparatus used there is no relationship between the angular position of the handwheel knob and the linear position of the cursor. The position of the cursor is solely dependent upon the rate of cranking. 3. The target remains stationary. 4. No changes in handwheel load occur with changes in the required angular rate of cranking. The "feel" of the handwheel does change when linear rate is changed for a constant angular speed. This results from the greater leverage of the larger handwheels.

#### Apparatus<sup>1</sup>

Figure 1 is a schematic diagram of the apparatus. The subject turns the crank with his right hand. The center of rotation is 81 cm. above the floor and the subject is seated with his right arm directly in line with the crank shaft. Friction in the system is kept at a low value by the use of ball bearings, and there is little inertia in the handwheel.

Attached to the end of the crank shaft is a small pulley that drives a second pulley mounted on the shaft of a tachometer genera-

tor. The output of the generator drives two mirror galvanometers that are wired in series with the generator. Separate light sources are focused on each of the mirrors. A patch of light 1 cm. high and 3 mm. wide is reflected from one of the mirrors to the back of a ground-glass screen that is perpendicular to the subject's line of sight. The light strikes the glass at a point 119 cm. above the floor. Two vertical black lines, 3 mm. apart, are drawn on the screen. These black lines serve as the target. The subject's task is to center the reflected patch of light between the two lines by turning the crank at a constant rate. Variation in the required angular rate of turning is achieved by adjusting a potentiometer that controls the resistance in the circuit between generator and galvanometers. The greater the resistance in the circuit, the higher the rate of cranking required to center the light-spot (cursor) between the target lines. A circular spot of light reflected from the second galvanometer to the surface of a photoelectric cell provides the main source of error indication in the apparatus. Error tolerance is determined by the angular position of the photocell relative to the recording

<sup>1</sup> The apparatus was constructed by Mr. Ervin G. Smith, Jr. of the Engineering Laboratory, Institute for Cooperative Research, The Johns Hopkins University.

galvanometer. When the subject turns the handwheel at the required rate, plus or minus the chosen error tolerance, the light from the recording galvanometer strikes the photocell. This activates an electric clock that is read to the nearest .01 of a second. Another clock times the trial period and shuts off the apparatus at any selected trial-length. The timing and recording clocks do not begin to operate until the subject first reaches the required rate of turning. Because of these features the number of seconds accumulated on the dial of the recording clock indicates the total time, during a trial period, in which the subject maintains a rate of turning within established tolerance limits.

Two other indications of performance are obtained simultaneously. One electric counter indicates the number of times during a trial period that the subject's rate of turning falls within the tolerance limits. This is a measure of the frequency of oscillation in rate within a trial. A second electric counter counts the number of revolutions actually turned during a trial. At low speeds accuracy in counting is achieved by counting to the nearest .2 of a revolution. At the higher speeds only one count per revolution is possible. Both the recording clock and the counters are enclosed in soundproofed boxes.

Variation in the radius of turning is achieved by attaching the knob to the handwheel at various distances from the center of rotation. A counter weight is also provided.

Error recording for counterclockwise turning is accomplished by reversing the tachometer generator end-for-end since the generator does not produce reliable voltages when the direction in which the shaft turns is reversed.

### Procedure

Only right-handed male subjects were used in the experiment. Fifteen subjects turned the handwheel in the clockwise direction, and fifteen different subjects turned the handwheel in the counterclockwise direction. All subjects received five consecutive trials, 30 seconds in length, at each of five angular rates of cranking combined with each of three crank radii. The rates used were 25, 75, 125, 175, and 225 revolutions per minute. The crank radii were 2.5, 7.5, and 12.5 cm.

The orders in which the subjects cranked under the various conditions were determined by a  $15 \times 15$  latin square. The same square was used for both directions of cranking.

Subjects were instructed to keep the cursor-light centered between the two target lines at all times. The apparatus was so arranged that the cursor moved to the right as the speed of turning increased from zero velocity for clockwise turning. For counterclockwise cranking the cursor moved to the left as speed increased. The range of error tolerance was kept at  $\pm 9\%$  of the required rate of turning. Subjects were not aware that there was any tolerance in the scoring system.

### Results

*Effects of Linear and Angular Rates.* In this experiment linear rate was varied independently of angular rate by changing the radius of movement for various constant angular speeds. When the angular speed of movement is constant, rate-accuracy increases with increased linear rate in the lower range of handwheel speeds. At the higher handwheel speeds this relationship is reversed, and accuracy decreases as linear rate increases.

Figure 2 pictures these results which are similar to those obtained by Helson with a compensatory tracking task (3). Helson, however, did not distinguish between angular and linear rates of movement. In Figure 2 linear rate increases with increased handwheel radius for any one handwheel speed, but the actual linear rates are not equal for the same radius at different handwheel speeds.

The significance of the differences between handwheel speeds, radii, and the interaction between these two variables was tested by the non-parametric analysis of variance described by Friedman (2) and Wilcoxon (6). This procedure was necessary because the data exhibited heterogeneity of variance when subjected to Bartlett's test (1). For both directions of turning the two main variables and the interaction between them were significant sources of variation ( $p < .01$ ). The results for the two directions of movement are combined in Figure 2 because they show similar tendencies.

Figure 3 is a graph of the relationship between accuracy and angular rate when linear



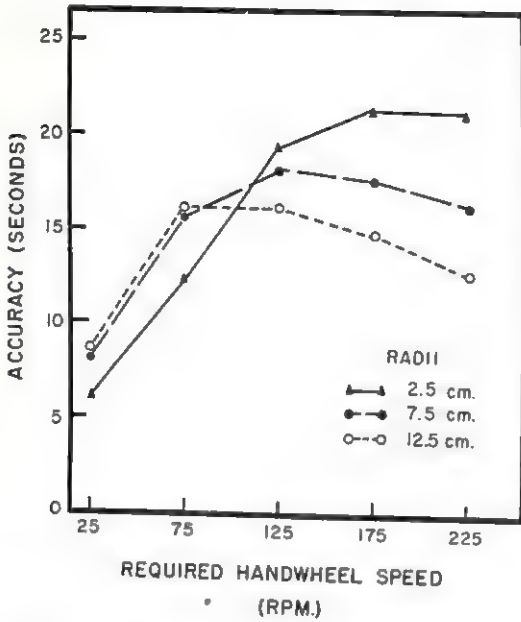


FIG. 2. Accuracy as a function of handwheel speed for different crank radii. The ordinate values indicate the mean time that the rates were maintained within specified tolerance limits. A score of thirty is the maximum possible accuracy score.

speed is constant. The values in the figure were obtained by interpolation between points on Figure 2 and calculation of the actual linear rates. Linear rate may be held constant while angular rate changes by adjusting the radius of the cranking movements.

Figure 3 shows that, for various constant linear rates, rate-accuracy increases with angular speed up to about 175 rpm. For the lower linear rates it appears that slightly greater accuracy might be obtained at speeds beyond 175 rpm. It would be necessary to extrapolate values in order to extend all curves over the entire range of angular speeds.

There is one factor which does not remain constant when linear speed is held at a fixed value by adjusting handwheel size as angular speed increases. Different muscles become involved in the control of the handwheel as size is changed. This factor is of relatively little importance for the handwheels used in this study.

It might be expected that increased accuracy would result from increased angular or linear rates of movement since, as Helson has pointed out (3), the absolute sensitivity of the handwheel decreases as rate increases.

For example, with a  $\pm 9\%$  error tolerance, the range of permissible error is about  $\pm 2.2$  rpm at a rate of 25 rpm, and  $\pm 20$  rpm at a rate of 225 rpm. Inspection of Figure 2, however, indicates that at the higher handwheel speeds little advantage is taken of the decreased sensitivity. For the larger radii accuracy actually drops off with handwheel speeds greater than 75–125 rpm. This effect cannot be related to the physical limitations of the subjects. Inspection of the records obtained with the revolutions counter showed that, with 11 minor exceptions in 450 trials, all subjects were capable of cranking with all radii at an average angular rate greater than 175 rpm when attempting to achieve the rate of 225 rpm.

The data concerning the frequency of oscillations in rate within a trial suggest that the oscillatory nature of cranking movements places a limit on the rate-accuracy achieved. These frequencies are shown in Figure 4 in which the data for the two directions of turning are again combined.

A single oscillation in rate, as measured in this experiment, includes both the change in

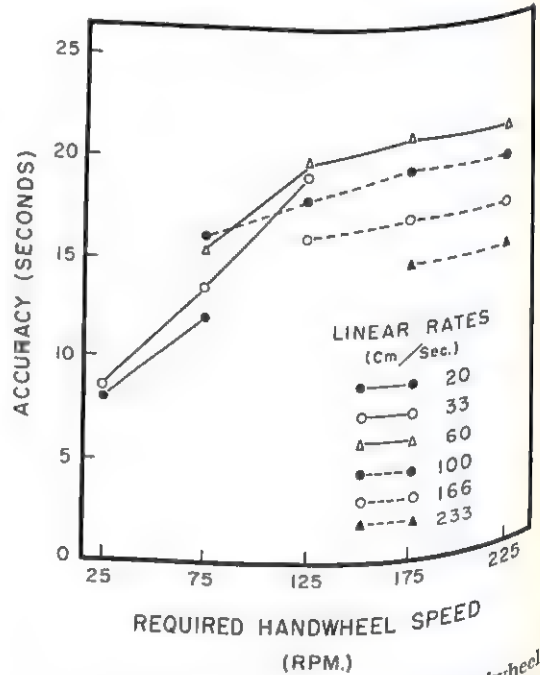


FIG. 3. Accuracy as a function of handwheel speed for different constant linear rates. The values on the curves were obtained by interpolation between points in Figure 2 and calculation of the linear rates.

rate greater than the error tolerance and the return to a rate within the error tolerance. From Figure 4 it is apparent that the number of oscillations increases with both linear and angular rate in spite of the reduction in sensitivity. For the two larger radii the number of oscillations increases at an increasing rate as the required handwheel speed is raised. In contrast, the number of oscillations for the smallest radius increases at a much slower rate.

Non-parametric analysis of variance established the significance of the over-all effect of handwheel speeds and radii upon the number of oscillations per minute ( $p < .001$ ).

Figure 5 shows the durations of the mean rate-oscillations in seconds for the various handwheel speeds and radii. The durations plotted in the figure were obtained from Figures 2 and 4. The accuracy scores for each point in Figure 2 were first subtracted from the maximum possible score of 30 seconds. The resulting values indicated the time spent outside of the error tolerance in an average trial. These scores were then divided by the number of oscillations in rate for the appropriate points shown in Figure 4. The obtained values were the mean durations of the rate-oscillations in seconds. Figure 5 shows

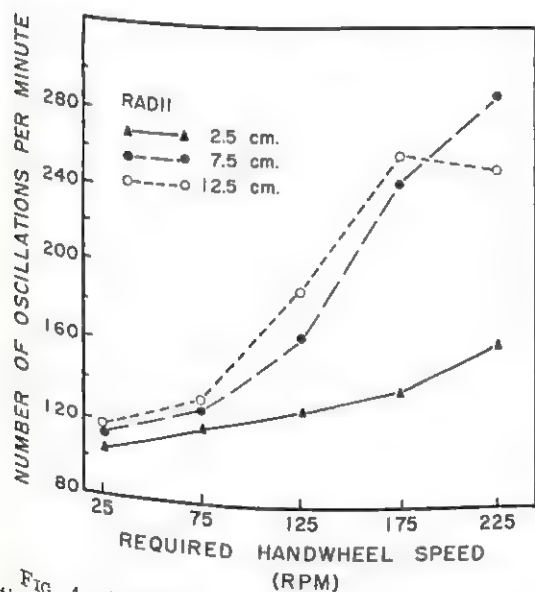


FIG. 4. Frequency of rate-oscillations as a function of handwheel speed for different crank radii. A greater oscillation includes both the change in rate greater than tolerance limits and the return to a rate within tolerance limits.

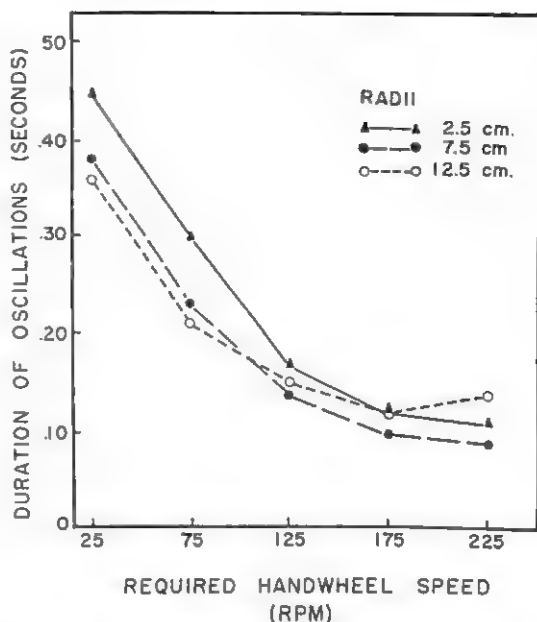


FIG. 5. Durations of mean rate-oscillations.

that the duration of the mean oscillations in rate is a decreasing function of handwheel speed.

Considered together, Figures 4 and 5 show why accuracy in Figure 2 does not continue to increase above certain speeds as the sensitivity of the handwheel decreases. At the lower handwheel speeds the number of rate-oscillations increases slowly as angular and linear speeds increase in combination, but the durations of the oscillations decrease rapidly. More errors are made, but they are eliminated much more quickly. The result is increased accuracy with increased speed. In the middle range of handwheel speeds the durations of the oscillations still decrease as speed increases, but at a much slower rate. At the same time, however, the number of oscillations is increasing rapidly. The result is a levelling-off and even a decrease in accuracy.

This interpretation also accounts for the relative accuracy obtained with different radii of movement. At the low handwheel speeds, for example, a greater number of oscillations appear with the larger handwheels. However, the durations of the oscillations are shorter for the larger handwheels and increased accuracy results.

Figure 5 provides suggestions concerning the nature of the responses involved in main-

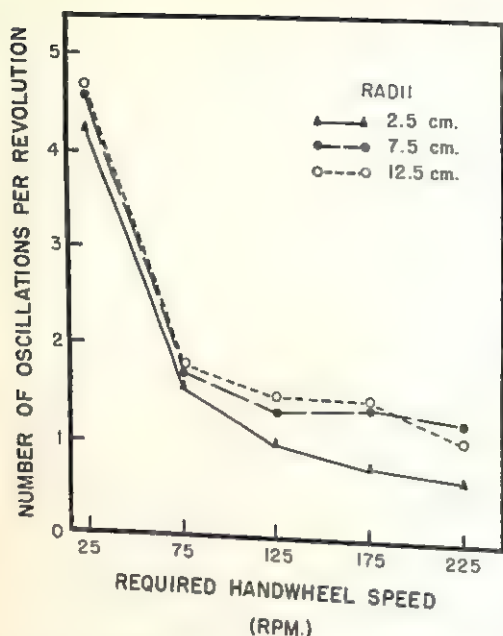


FIG. 6. Number of rate-oscillations per handwheel revolution.

taining a rate of movement with a handwheel. The durations of the average oscillations are too short to have allowed the initiation of corrective responses on the part of the subjects. In having access to the instantaneous indication of rate provided by the apparatus, subjects received more information than they could use. Apparently subjects accepted the unsteadiness in their movements as being beyond their control and tried to center their oscillations about the target in order to minimize error. These results support the suggestion of Lincoln and Smith (5) that accuracy in direct pursuit-tracking with a handwheel control is limited by the oscillatory nature of the tracker's response.

Figure 6 indicates the number of oscillations in rate per handwheel revolution. At the higher handwheel speeds the relative number of oscillations is fairly constant for the different radii. The actual number approaches a rate of one oscillation per revolution. This may mean that there is a more pronounced tendency for deviations in rate to occur when the handwheel knob is in a particular spatial position. The most likely position would be at the top or bottom of the swing during rotation.

*Effects of Direction of Cranking.* Although the accuracy achieved in clockwise turning was slightly higher than for counterclockwise turning on four of the five speeds, these differences were not significant when tested by the necessary non-parametric methods. Differences between directions were established, however, in another measure of performance—the constant rate error.

Figures 7 and 8 show the constant errors in the average rate of cranking in rpm. Figure 7 is for clockwise turning, while Figure 8 is a plot of the data for counterclockwise turning. For both directions the sign of the constant errors indicates that, considered as groups, the subjects tended to crank at rates that were slower than the required rate, although they were capable of turning at higher rates. Beyond the speed of 175 rpm this statement does not hold since some subjects could not maintain the rate of 225 rpm.

The constant errors for the counterclockwise direction are significantly greater than for the clockwise direction ( $p < .01$ ). In addition, constant error shows a significant increase ( $p < .01$ ) in size as linear rate increases with constant angular rates for both directions of cranking. For the counterclockwise direction only, constant error shows a significant increase ( $p < .001$ ) in size as angular and linear rates increase in combination. The speed of 225 rpm was not included in the calculation of the latter two probabilities.

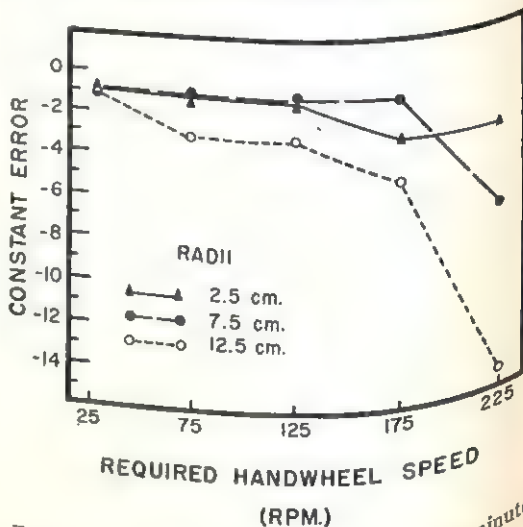


FIG. 7. Constant error in revolutions per minute for clockwise turning.



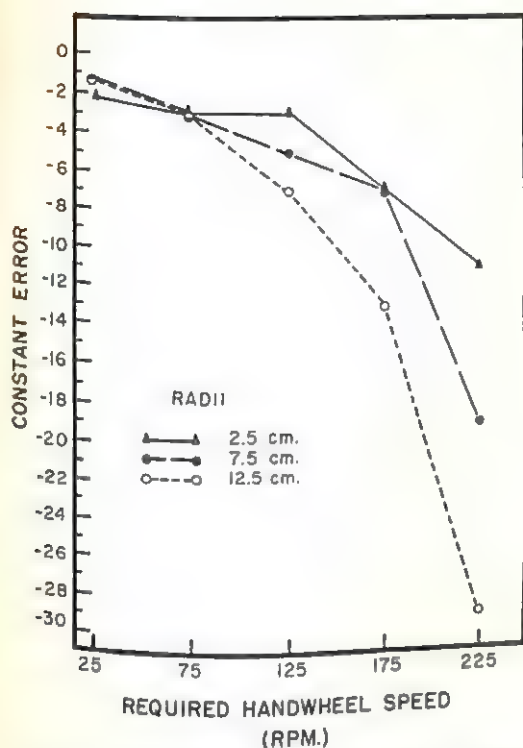


FIG. 8. Constant error in revolutions per minute for counterclockwise turning.

because the error at that speed is influenced by the physical limitations of the subjects.

If subjects do tend to ignore their unsteadiness in rate and center their oscillations about the target, they do so with a persistent bias toward rates that are slower than the required rate. The bias is greater for cranking in the counterclockwise direction.

### Summary

Subjects cranked a handwheel at each of five different speeds combined with each of three different handwheel radii. Fifteen subjects cranked in the clockwise direction while fifteen other subjects cranked in the counterclockwise direction.

The subjects were provided with an instantaneous visual indication of their rate of cranking. In appearance the task resembled a conventional tracking problem. With the apparatus used, however, the task was reduced to the maintenance of the required rates since positional relationships between the rate indicator and the handwheel knob were eliminated, and it was impossible for the

subjects to achieve the required rate without aligning the indicator and target. In addition, no changes in handwheel load were introduced by changes in the required speed of turning.

At the lower handwheel speeds, rate-accuracy improved with increases in the linear rate of movement for a constant angular rate. At the higher angular speeds an inverse relationship appeared between linear rate and accuracy. Linear rate refers to the units of distance traveled per unit of time by the handwheel knob. Linear rate was varied independently of angular rate by changing the radius of the movement.

For constant linear rates accuracy always improved with increased angular rate up to about 175 rpm. The failure of accuracy to continue to improve above a certain point when linear and angular rates were increased in combination was attributed to the oscillatory nature of cranking movements.

Subjects tended to crank at rates slower than the required rate although they were capable of maintaining the required rate. This tendency increased as both linear and angular rate increased. No significant differences in accuracy appeared between the two directions of movement, but those subjects who cranked in the counterclockwise direction showed a significantly greater tendency to lag in rate.

Received November 16, 1953.  
Early publication.

### References

1. Edwards, A. L. *Experimental design in psychological research*. New York: Rinehart, 1950.
2. Friedman, M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J. Amer. statist. Ass.*, 1937, 32, 675-701.
3. Helson, H. Design of equipment and optimal human operation. *Amer. J. Psychol.*, 1949, 62, 473-497.
4. Lincoln, R. S. and Smith, K. U. Systematic analysis of factors determining accuracy in visual tracking. *Science*, 1952, 116, 183-187.
5. Lincoln, R. S. and Smith, K. U. Visual tracking: II. Effects of brightness and width of target. *J. appl. Psychol.*, 1952, 36, 417-421.
6. Wilcoxon, F. *Some rapid approximate statistical procedures*. American Cyanamid Co., 1949.

## Applied Psychology in Action

### Reply to Dr. Wells and to Miss Epstein \*

Howard D. Hadley

*Morey, Humm, and Johnstone, Inc., New York, N. Y.*

Dr. Wells is correct in his analysis. I should have used credulity.

I should also like to make some comments about Miss Epstein's note. In therapy, you are concerned, at least at first, with reducing threats. In advertising, where you are dealing with more "normal" persons, the task at hand is to offer enhancements to the consumer. In this latter case, threat is something to be avoided, not to be "cured." In both instances, a sympathetic atmosphere is a primary requisite. I was mainly concerned with the *atmosphere* created by both advertisers and therapists.

Miss Epstein is somewhat correct when she says that the non-interference principle is not

\* Wells, F. L. Comment on word meaning. *J. appl. Psychol.*, 1954, 38, 133. Epstein, Mary. A note on "the non-directive approach in advertising appeals." *J. appl. Psychol.*, 1954, 38, 133-134.

applicable to advertising. If there were no "interference," there would be no selling. However, doesn't a patient have an attitude towards the therapist at the end of the sessions? Also, aren't these attitudes often favorable? By not "interfering," may not a favorable attitude be developed?

Actually, there is no pure example of inferred advertising. Direct and inferred were contrasted to sharpen the concept. Just as there are few (if any) completely introverted or extroverted persons, there are few (if any) advertisements which are completely direct or inferred.

In the end, it is the atmosphere created by the advertisement that is important. Direct-inferred, directive-nondirective, are more logical constructs than useful tools. It's the end result that is important.

## Note on the Work of the British Standards Institution

K. F. H. Murrell

*Department of Ergonomics, T. I. (Group Services) Limited, Birmingham, England*

In 1949, when I was Head of the Naval Motion Study Unit in the British Admiralty, I was invited to serve as an Admiralty representative on a British Standards Institution Committee, which was considering the design of pressure gauges. I had for some time been studying published research of workers such as C. J. Berger, A. Chapanis, W. F. Grether, W. R. Garner, W. E. Kappauf, R. B. Loucks, R. E. Sleight, S. D. S. Spragg, and M. J. Warwick on factors influencing the readability of dial faces, and I advised this Committee of these findings. Most of them differed too radically from existing practice

for them to be acceptable to the trade as standards but some were adopted mainly by way of footnotes and recommendations.

The British Standards Institution subsequently decided to set up a Technical Committee to produce a "code of practice" on the graduation and marking of instruments, using as its starting point an Admiralty Report (Naval Motion Study Report No. 48) which I had written summarizing all available research on the subject. This Committee has now been sitting for two years and not only has it considered the findings of existing research but it has also arranged for further re-

search to be done to fill gaps in our knowledge. One such experiment is being carried out at the moment on the relationship between dial size and accuracy of reading.

This is, I believe, the first instance of psychological research being used as the basis of deliberations by the British Standards Institution.

THE JOURNAL OF APPLIED PSYCHOLOGY  
Vol. 38, No. 3, 1954

## Personnel Psychology and Small Business

W. Grant Dahlstrom

University of North Carolina

The psychologist acting as consultant to a small concern operates under serious restrictions on his usual personnel methods. Often the selection of employees is only a small part of his general services. The number of new men being hired is small, the turnover rate may be infinitesimal, and even a survey of the men on the job may yield paltry information in comparison with usual employment studies. Ideally this sort of operation should be carried on at a national level with many or all of the local concerns of the same sort participating in the project. This would probably only be feasible through arrangements between an association of these concerns and some psychological consultants similar in scope to the Psychological Corporation. Lacking these connections, the psychologist locally still faces these persistent difficulties.

When the local concern invests considerable time and money in job training a professional level man, the contribution of the psychological consultant in screening applicants could be valuable. The method of choice in handling this problem would include the utilization of screening tests. But the situation is quite frustrating because of the difficulties in the way of establishing score standards on these tests. In the opinion of the writer, there are three lines of evidence which can lend support to the consultant in making his judgments and recommendations.

One of these is the degree of homogeneity in the test data provided by a survey of existing employees. Recently the writer obtained such data on a firm of medical consultants

providing a business advisory service to physicians and dentists. There were only 10 field men and 2 central office men who had previously acted as medical consultants themselves. The results on the *Strong Vocational Interest Blank* gave a compelling impression of homogeneity. (In the group, 5 men had A ratings in both areas VIII and IX; 4 in IX; and 3 others in VIII. The only other A ratings occurred in area III twice, area V twice, and areas VII and X once each. Very low ratings appeared in area II in more than half the group.) The *MMPI* results, though less uniform, were also rather homogeneous. (The triad of scales Pd, Pa, Ma were at a codable level (above 54 T-score) in half the group, with Pa appearing at this level in 10 of the 12 tests. The K scores ranged above 60 T-score without exception.)

Obviously such a concept of homogeneity is relative. The men are very similar when you consider the assorted patterns from men-in-general. The uniformity is less impressive if reference is made to male college graduates only, or to business majors, or even more appropriately to those men with sufficient training to be considered at all for employment in such a company. Nevertheless, this seems a workable concept. If sufficient data were available in a usable form on various general groups, the psychologist could make a judgment about the homogeneity resulting from selective survival. Not a great deal of the normative data on our multiscale tests is published in a form in which we can judge relative frequency of particular score combina-



tions. A quantitative index of relative uniformity could be devised which would serve better than judgment in this matter.

This is not meant to imply that experience on the job, faking of test responses, and similar sources of variance may not also be operating to produce this uniformity, but as research information is accumulated on such factors as these, reasonable allowance could be made for them as well. This is one of the most workable meanings that can be offered for *theory*, as it is to be used in the areas of employee selection or vocational choice as discussed by Dr. Super in his divisional presidential address.

The second line of evidence available to the psychologist involves precisely this matter of theory. If the findings from the test survey are not only satisfactorily homogeneous, but in addition conform to expectations based on considerations of the job components, then the consultant can feel on even firmer ground in making his decisions.

In this particular instance, talks with the psychologically shrewd assistant manager in the company, as well as informal interactions with Dr. Clayton Gerken at the State University of Iowa, led the writer to formulate certain characteristics of the *ideal* medical consultant. These speculations involved a marked interest in business detail combined with an equally strong interest in business contact. It was also expected that an executive interest level and maturity would show up on the tests. This actually proved to be the common core of test results from the group. Workers more familiar with the tests

and with a wider knowledge of job breakdowns could erect a much more detailed set of expectations. Psychological consultation will always involve a modicum of this sort of psychologizing, even in the face of greater usage of actuarial methods at specific decision points. Accumulated research findings should facilitate the formulation of these working hypotheses.

The third line of evidence stems from the correspondence between a man's rated proficiency on the job and the degree of approximation in test score pattern to the *ideal* employee. Here restriction in range of score dispersion is a serious limitation, and if the psychologist finds even a moderate relationship, he may assume more satisfactory values would be obtained with a wider sampling.

This last point is obviously the most deceptive line of evidence of the three since such eventualities as non-rectilinearity in the correlation surface, unexpected discontinuities in the functions, or errors in the validity data may arise to embarrass him. These assumptions would have to be continually checked against research findings from psychologists more favorably situated in respect to criterion data and research samples.

If the consultant takes the trouble to institute a testing program and finds relative homogeneity, consistency with expectations, and fair corroboration of test findings with on-the-job performance, he can operate with considerably more confidence and effectiveness on even small projects than he could on the basis of sheer intuitive speculation alone.

## Book Reviews

Kinsey, A. C., Pomeroy, W. B., Martin, C. E., Gebhard, P. H., et al. *Sexual behavior in the human female*. Philadelphia: W. B. Saunders Company, 1953. Pp. xxx + 842. \$8.00.

Hiltner, S. *Sex ethics and the Kinsey reports*. New York: Association Press, 1953. Pp. xi + 238. \$3.00.

Aberle, Sophie D. and Corner, G. W. *Twenty-five years of sex research. History of the National Research Council Committee for Research in Problems of Sex, 1922-1947*. Philadelphia: W. B. Saunders Company, 1953. Pp. v + 248. \$4.00.

Applied psychology has grown in direct ratio to the accumulation of facts. But in the area of sex behavior, answers to the question (with apologies to *Dragnet*) "What are the facts, Ma'am?" have been hard to come by. This is due chiefly to our puritanical heritage which has thrown such strong taboos around the subject. In a real sense, we are now witnessing the final retreat of the censor in denying man the right to knowledge of the natural functions of the body. This retreat began 400 years ago when Vesalius dared to pierce some of the mystery of what goes on inside the human body and, 100 years later, his disciple, William Harvey discovered circulation of the blood.

One hesitates to comment in detail concerning the epoch-making work of Kinsey and his co-workers. This is so because of the extensive reviews that have already appeared in connection with his earlier publication on the male. It is also because his book on the female has already received such widespread publicity in newspapers, magazines, and on radio and TV. One would only be repeating what so many critical and uncritical evaluators have already written or said. For this reason, this review merely points to the facts that have been marshalled in the enormous range of individual differences in reported capacity, or rather claimed performance, and to the extensive detailed information set forth in Part III entitled "Comparisons of Female and Male." The sex difference shown by the

fact that in 30 out of 33 items the male, on the average, is more readily affected by psychological stimuli is worthy of special note. This finding provides a wealth of insight for better understanding of the psychology of human males and females.

The applied psychologist would do well to follow closely discussions of Kinsey's work by representatives of organized religion. Reverend Seward Hiltner, who is pastoral consultant to the Editorial Advisory Board of *Pastoral Psychology Magazine*, which was founded in 1950, has written a detailed interpretation of all aspects of Kinsey's reports. His book will aid many clergymen to assimilate the findings with a minimum of trauma. It should enable them to do a better job of understanding and counseling their parishoners, young or old. Enlightened premarital and marriage counseling as a pastoral duty has been going on to an ever-increasing extent for a generation. Hiltner's book will undoubtedly accelerate this important movement.

The significance of Kinsey's work and of Hiltner's interpretation can be fully understood only by studying the magnificent achievements of the NRC Committee for Research in Problems of Sex. Aberle and Corner's report gives due credit to Robert M. Yerkes who was chairman from 1922 to 1947. Yerkes' foresight, initiative, tact, courage, and everlasting persistence were primarily responsible for this development. He was able to secure the collaboration of top-notch scientists. He was also able to secure continuity of financial support. The Committee courageously supported research on all aspects of sex in all species from paramecium to man. Scores of researches were supported and hundreds of research reports were published in a wide range of scientific journals, monographs, and books. The bulk of the work was directed toward infrahumans but, from the beginning, research on sex behavior in man was strongly supported. The latter studies were begun by R. S. Lee, Adolph Meyer, L. M. Terman, and W. R. Miles in the nineteen twenties, were continued in the

thirties by Carney Landis, E. Lowell Kelly, and Terman. Since 1937, Dr. Kinsey and his group at Indiana University have been the chief beneficiary.

Thus psychology has moved from an intellectualistic preoccupation with man as a rational being to a more realistic understanding of man as a behaving organism in *all* of his manifold adjustments. In short, sex can no longer be ignored.

Donald G. Paterson

University of Minnesota

Lundin, R. W. *An objective psychology of music*. New York: Ronald Press, 1953. Pp. ix + 303. \$4.50.

This book is a noteworthy addition to the psychology of music, especially for classroom use with the undergraduate student. Its style is clear and simple, its coverage is unusually comprehensive, and its range is wide. It will truly facilitate the learning process for the student, an advantage which has often been lacking in this field. The psychology of music demands an understanding of two very different disciplines, one of them a science, the other an art. The vocabulary and style employed by the artist has often proved baffling to the scientist, and vice versa. Lundin has shown a special talent as an interpreter, and has made his material thoroughly clear to both. His occasional oversimplifications will prove justifiable in terms of the student who seeks competency in two fields.

In the history of the psychology of music the striking advance represented by the Sea-shore tests has actually turned out to be one of the greatest disadvantages. Preoccupation with this temptingly simple but peculiarly inadequate instrument has served as a block to the development of better measuring scales, and the imitators of the great pioneer have done little to improve on his original work. Lundin has thrown all these studies into better perspective and his summaries and evaluations should save many future missteps in this special field of investigation.

The most significant guideposts in this particular area of musical research seem to point in the direction of cultural rather than nativistic explanations and interpretations.

Lundin does well therefore in taking a stand for an interbehavioral point of view, and steering his readers away from the unsubstantial rhapsodies of Howe and the even more refined semantics of later writers, toward the more substantial investigations and arguments, supported by something more than fine writing. This interbehavioristic trend which Lundin endorses so emphatically pervades the thinking of many writers even though they identify themselves with less exacting and more eclectic schools. Farnsworth, one of its strongest advocates, has made this point of view much more attractive and stimulating by demonstrating its usefulness in a varied program of research.

From many quarters there are evidences of renewed interest and activity in the problems of esthetics. Audiences and amateur performers in both the graphic and theater arts, but especially in music, are in a period of expansion. Now that the groundwork has been so carefully laid, any pedagogue or experimentalist, and even Lundin himself, can go on to be more persuasive and more fruitful in developing the psychology of music. The whole field has been brought up to date, the arguments are sound, and the movement toward further knowledge has been greatly accelerated by this timely and much needed book.

(Mrs.) Kate Hevner Mueller

Indiana University

Woolf, M. D. and Woolf, Jeanne A. *The student personnel program*. New York: McGraw-Hill Co., 1953. Pp. ix + 416. \$5.00.

Subtitled, *Its development and integration in the high school and college*, this book is an "... attempt to picture a comprehensive student personnel program ..." (p. v) and draws heavily on the authors' twenty-six years of educational experience. Most professional workers will find much of interest in it. Most will also be bothered by a number of shortcomings.

Virtually all phases of the student personnel program are dealt with. After a short introductory chapter on the expanding role of the personnel worker, there are chapters



on counseling, group methods, student government, discipline, housing, remedial services, measurement, orientation, faculty advising, training of personnel workers, and administration of the program. Unfortunately, there is no readily discernible framework, no integrating philosophy which might have given added meaning to the extensive content. The chapters seem almost to be separate essays, and the demands on the reader are not always rewarded.

It is difficult to specify a single group for which this book is entirely appropriate. Professional workers will find parts of it quite elementary although the many examples will be interesting. Beginning students will not have the necessary backgrounds of information to give it the critical reading it requires. Academic administrators are likely to become bogged down in details which are not woven into a meaningful fabric.

It is an unnecessarily difficult book to read, perhaps because the authors appear not to have been sure whether they were doing a primarily scholarly work or one based mostly on their own experiences, comments in the preface notwithstanding. Those sections reporting their experiences are probably the best in the book. The scholarly sections are cursory and not well done. Questionable scholarship is indicated, for example, by an acknowledgment to another staff counselor for pointing out that "... the refusal of the counselor to let the client lean on him may be actual rejection" (p. 33). The theories of misbehavior discussed in the chapter on discipline do not represent the best of modern psychology. Elsewhere, they say, without reference to research, "If on the American Council on Education Psychological Examination, the L score . . . percentile ranking is twice that of the Q score . . . or vice versa, there is uneven mental development and often a personality adjustment problem" (p. 227).

The chapter on counseling is uneven, and the Chicago point-of-view is given such emphasis as to suggest a pre-eminence not yet granted by most counselors. The chapter on faculty advising is excellent for its many practical suggestions but is marred by a poorly handled survey of the literature. The

authors should be commended for their forthright discussion of training requirements in which psychology is placed at the center of the program.

Taken as a whole, the defects of this book seem to stem from two principal sources. In the first place, there is the previously mentioned apparent confusion about whether this is a scholarly book or one primarily reporting on experiences. Both are valuable and necessary, but they need to be carefully amalgamated, not cut and patched together. Secondly, there is the lack of a carefully thought out and explicitly stated philosophy of student personnel work. The tissues of a good book are presented, but there is no articulating skeleton.

John W. Gustad

University Counseling Center,  
University of Maryland

Leitner, K. *Hypnotism for professionals*. New York: Stravan Publishers, 1953. Pp. 127. \$4.00.

Konradi Leitner was a stage hypnotist who became quite well known through his work with the USO during the war. In this post-humorous book he describes his methods of working before an audience. He does this in a clear and interesting manner but he has contributed nothing to the scientific knowledge of hypnosis. There are a number of illustrations in the book using a very pretty feminine model with whom I'm sure any hypnotist would be happy to work.

William T. Heron

University of Minnesota

Jennings, E. E. *Techniques of successful foremanship*. Madison: School of Commerce, Bureau of Business Research and Service, University of Wisconsin, Wisconsin Commerce Studies, Vol. I, No. 4, March, 1953. Pp. 41. \$1.15.

This is the report of a study undertaken for the purpose of gaining an understanding of the techniques or traits characteristic of successful foremen prior to undertaking a supervisory training program. The "Jennings Supervisory Analysis," a 23-item questionnaire, was administered to 1,682 workers and

their 52 foremen in a large midwestern plant. All workers filled out the questionnaire by checking items which "outstandingly" described their own foremen. Every third worker filled out a second questionnaire, checking the 3 items he considered most desirable in foremen. The 52 foremen filled out 2 questionnaires. On the first, they checked items which best described their own behavior; on the second, they checked the 3 items they considered most desirable in foremen. Foremen were rated for over-all ability by pooling ratings of their immediate superiors with those made by top management in the plant. An appendix describes the method used in obtaining and pooling these ratings.

Findings presented are relative to the 23 items in the questionnaire. This, unfortunately, places a limitation on their meaning because all 23 items are favorable characteristics, selected when the questionnaire was developed on the basis of being "both generally descriptive and desirable" of foremen. With this limitation, findings include the following: (1) the 3 most desirable techniques of foremen are to be fair to everyone, go to bat for workers, and give clear-cut instructions; (2) there is little relationship between what workers feel is desirable in foremen and their descriptions of their own foremen; (3) there is little relationship between what workers and foremen feel is descriptive of foremen; (4) foremen and workers largely agree on what the desirable characteristics of a foreman are; (5) traits or characteristics descriptive of foremen rated as successful by their superiors are also considered desirable by workers.

There is no indication in the report of the audience for which it is intended. It appears, however, that it was not intended for persons with technical background in that much of what should be included in a report for this group is lacking. Only gross rankings are presented, without averages or measures of variability. Correlation coefficients are used without descriptions of the method of computation used. Interpretations of statistical findings are also questionable in some cases. For example, in an item analysis in

which frequency with which a foreman was described by an item was correlated with success as indicated by superiors' ratings, 12 correlations, ranging from .28 to .53, are presented as evidence that these 12 items are "highly related" to success. The discussion sections of the report go beyond the data presented, although the author does point out that intensive interviewing of foremen was an additional source of information.

Although this report presents objective evidence on desirable foreman characteristics, it is doubtful whether the author's hope that the findings "... can be used to both clarify the objectives and to increase the effectiveness of foreman training programs" will be realized by persons who turn to this report with that same hope in mind.

Theodore R. Lindbom

*Midland Cooperatives, Inc.,  
Minneapolis, Minnesota*

Montagu, A. *The natural superiority of women*. New York: Macmillan, 1953. Pp. 205. \$3.50.

The first question that confronts the reviewer in evaluating this book is "Just why was it written?" The facts brought out here about sex differences have been available for some time to intelligent men and women like the readers of the *Saturday Review* for whom this presentation was first designed. The books of Amram Scheinfeld and Margaret Mead on the subject have been widely read.

It seems that this particular work is a polemic rather than simply a popularization of scientific material. Ashley Montagu, like other writers of our time, is concerned about the state of our society. Many of our difficulties arise, he thinks, because of the emphasis we place on aggressiveness and competition and our failure to promote loving-kindness and cooperation. These policies he sees as a consequence of the long-continued subjection of women. The psychological characteristics in which they excel have been systematically devaluated and thus both men and women have failed to stress the values which alone can insure the survival of humanity. If we can become convinced that the female sex is biologically and psychologically

*superior*, that endurance and resistance to disease are more important than muscular size and strength, and that emotional expressiveness and social perceptiveness are more important than aggressiveness and mechanical skill, we shall have taken the first step toward the new emphasis our times require.

The necessity the author sees to make the case that women are in all ways superior is responsible for the book's major defects. In the first place, he insists again and again that what he is about to say will be shocking to the reader. One reads on to encounter some rather innocuous fact such as the difference in survival rates or frequency of automobile accidents. Secondly, his argument often leads him into reasoning which has sometimes been unkindly labeled "feminine" logic. In Chapter 4, for example, he explains first that there is no relationship between brain weight and intelligence and then goes on to argue that women's brains are actually larger than men's in proportion to total body size. If brain weight is a matter of no importance, why insist on superiority with regard to it? Thirdly, the debating orientation produces a certain distortion in some of the facts themselves. On the topic of intelligence differences, for example, he devotes so much more space to *listing* all the kinds of evidence for superior verbal ability in girls than he does to summarizing the kinds of test material on which

boys excel, that his conclusion that girls do better than boys on intelligence tests, with a few insignificant exceptions, appears plausible. On page 121 he quotes Stoddard's statement as to the impossibility of evaluating sex differences in intelligence using our present tests, but by this time he has already *used* the available data to support his argument for female superiority. (Incidentally this is one of the facts he expects to be "shocking" to us.)

Few of us would quarrel with Mr. Montagu's desire to see more love and cooperation in our society or his conviction that good relationships between the sexes are vital. The question is how much such an approach as this contributes to these ends. On page 185 he asks, "Is it too much to hope that the claims herein made for the natural superiority of women will shake men out of their complacent acceptance of the present position of the sexes?" My answer would be, "Yes, I am afraid it is too much to hope. People's convictions are not that easily shaken." But whatever it may be worth as argument, for serious students of differential psychology, the contribution made by this book to our factual knowledge of sex differences can safely be ignored. There is nothing new here.

Leona E. Tyler

*University of Oregon*



## New Books, Monographs, and Pamphlets

Books, monographs, and pamphlets for listing and possible review should be sent to Donald G. Paterson, Editor, Department of Psychology, University of Minnesota, Minneapolis 14, Minnesota.

- Twenty-five years of sex research history of the National Research Council Committee for Research in Problems of Sex, 1922-1947.* Sophie D. Aberle and George W. Corner. Philadelphia: W. B. Saunders Company, 1953. Pp. 248. \$4.00.
- The nature of prejudice.* Gordon W. Allport. Cambridge, Mass.: Addison-Wesley Publishing Company, Inc., 1954. Pp. 544. \$5.50.
- Student personnel services in higher education.* Dugald S. Arbuckle. New York: McGraw-Hill Book Company, 1953. Pp. 352. \$4.75.
- The human person.* Magda B. Arnold, J. A. Gasson, et al. New York: The Ronald Press Company, 1954. Pp. 585.
- Educational psychology.* Glenn Myers Blair, R. Stewart Jones, and Ray H. Simpson. New York: The Macmillan Company, 1954. Pp. 601. \$4.75.
- Introduction to advertising.* Arthur J. Brewster, H. H. Palmer, and Robert G. Ingraham. New York: McGraw-Hill Book Company, 1954. Pp. 480. \$5.50.
- Handbook of probability and statistics with tables.* Richard S. Burington and Donald C. May, Jr. Sandusky, Ohio: Handbook Publishers, Inc., 1953. Pp. 340. \$4.50.
- Practical applications of democratic administration.* Clyde M. Campbell. New York: Harper & Brothers, 1952. Pp. 325. \$3.00.
- Studies in the scope and method of "The Authoritarian Personality."* Richard Christie and Marie Jahoda, Editors. Glencoe, Ill.: The Free Press, 1954. Pp. 279. \$4.50.
- Journal of personnel administration and industrial relations.* Edited by Lee W. Cozan. Vol. 1, No. 1, January 1954. Quarterly. \$6.00 per year.
- Rehabilitation of the older worker.* Wilma Donahue, James Rae, Jr., and Roger B. Berry. Ann Arbor: University of Michigan Press, 1953. Pp. 200. \$3.25.
- Building up the supervisor's job.* M. J. Dooher, Editor. New York: American Management Association, 1953. Pp. 35.
- Essentials of effective administration.* M. J. Dooher, Editor. New York: American Management Association, 1953. Pp. 51.
- Motivation: the core of management.* M. J. Dooher, Editor. New York: American Management Association, 1953. Pp. 44.
- The historical roots of learning process.* Horace B. English. New York: Doubleday and Company, Inc., 1954. Pp. 21. \$65.
- How to choose that career.* S. Norman Feingold. Cambridge, Mass.: Bellman Publishing Company, 1954. Pp. 52. \$1.00.
- Child development.* Ilse Forest. New York: McGraw-Hill Book Company, 1954. Pp. 286. \$4.00.
- The rating of performance with the help of films.* Paul F. Fornallaz. Madison: University of Wisconsin, School of Commerce, Bureau of Business Research and Service, 1954. Pp. 35. \$1.15.
- Feelings and emotions.* Lawrence K. Frank. New York: Doubleday and Company, Inc., 1954. Pp. 38. \$.85.
- Elements of statistics.* H. C. Fryer. New York: John Wiley & Sons, Inc., 1954. Pp. 263. \$4.75.
- Psychology applied to human affairs.* Second Edition. J. Stanley Gray. McGraw-Hill Book Company, Inc., 1954. Pp. 581. \$6.00.
- Measurement and evaluation in the secondary school.* Second Edition. Harry A. Greene, Albert N. Jorgensen, and J. Raymond Gerberich. New York: Longmans, Green and Co., Inc., 1954. Pp. 690. \$5.00.
- Teaching success of Catholic elementary school teachers.* Sister M. Mynette Gross. Washington, D. C.: The Catholic University of America Press, 1953. Pp. 129.
- Some observations on executive retirement.* Harold R. Hall. Boston: Harvard Business School, Division of Research, 1953. Pp. 298. \$3.75.
- How to lie with statistics.* Darrell Huff. New York: W. W. Norton & Company, Inc., 1954. Pp. 142. \$2.95.

- The process of psychotherapy.* Harrington V. Ingham and Lenore R. Love. New York: McGraw-Hill Book Company, 1954. Pp. 270. \$5.00.
- Improving supervisory behavior.* Eugene E. Jennings. Madison: University of Wisconsin, School of Commerce, Bureau of Business Research and Service, 1954. Pp. 35. \$1.15.
- P.I. Merit Rating Series.* Joseph E. King and Judith W. Wingert. Chicago: Industrial Psychology, Inc., 1953. Pp. 52. \$3.00.
- Developments in the Rorschach technique. Volume I: Technique and theory.* Bruno Klopfer, Mary D. Ainsworth, Walter G. Klopfer, and Robert R. Holt. Yonkers-on-Hudson, N. Y.: World Book Company, Publishers, 1954. Pp. 726.
- Statistical methods in experimentation: an introduction.* Oliver L. Lacey. New York: The Macmillan Company, 1953. Pp. 249. \$4.50.
- Better business communications.* Spencer A. Larsen, Editor. Detroit: Wayne University Press, 1952. Pp. 282. \$1.75.
- The natural man.* Clarence Leuba. New York: Doubleday and Company, Inc., 1954. Pp. 70. \$95.
- Measuring group cohesiveness.* Lester M. Libo. Ann Arbor: University of Michigan Press, 1953. Pp. 111. \$2.00.
- Handbook of social psychology. Volume I.* Gardner Lindzey, Editor. Cambridge, Mass.: Addison-Wesley Publishing Company, Inc., 1954. Pp. 704. \$8.50.
- Handbook of social psychology. Volume II.* Gardner Lindzey, Editor. Cambridge, Mass.: Addison-Wesley Publishing Company, Inc., 1954. Pp. 704. \$8.50.
- Educating the sub-normal child.* Frances Lloyd. New York: Philosophical Library, 1953. Pp. 148. \$3.75.
- Student counseling in Japan.* Wesley P. Lloyd. Minneapolis: University of Minnesota Press, 1953. Pp. 204. \$4.00.
- Areas of psychology.* F. L. Marcuse, Editor. New York: Harper & Brothers, 1954. Pp. 532. \$5.00.
- Understanding the Japanese mind.* James Clark Moloney. New York: Philosophical Library, 1954. Pp. 252. \$3.50.
- Revised Minnesota Occupational Rating Scales.* Donald G. Paterson, C. D'A. Gerken, and Milton E. Hahn. Minneapolis: University of Minnesota Press, 1953. Pp. 85. \$2.00.
- An introduction to clinical psychology.* Second Edition. L. A. Pennington and Irwin A. Berg, Editors. New York: The Ronald Press Company, 1954.
- Counseling: theory and practice.* Harold B. Pepinsky and Pauline N. Pepinsky. New York: The Ronald Press Company, 1954. Pp. 328. \$4.50.
- Music therapy.* Edward Podolsky. New York: Philosophical Library, 1954. Pp. 335. \$6.00.
- Mid-century crime in our culture.* Austin L. Porterfield and Robert H. Talbert. Fort Worth: Leo Potishman Foundation, Texas Christian University, 1954. Pp. 113. \$2.25.
- The personnel administrator at the crossroads.* John Post. New York: American Management Association, 1953. Pp. 54. \$1.25.
- Introduction to educational psychology.* H. H. Remmers, Einar R. Ryden, and Clellen L. Morgan. New York: Harper & Brothers, 1954. Pp. 420. \$4.00.
- Introduction to opinion and attitude measurement.* H. H. Remmers. New York: Harper & Brothers, 1954. Pp. 437. \$5.00.
- The high school student.* John W. M. Rothney. New York: The Dryden Press, 1953. Pp. 271. \$1.90.
- Personality dynamics.* Bert R. Sappenfield. New York: Alfred A. Knopf, Inc., 1954. Pp. 412. \$5.50.
- Psychological problems in mental deficiency.* Second Edition. Seymour B. Sarason. New York: Harper & Brothers, 1953. Pp. 402. \$5.00.
- The clinical interaction: with special reference to the Rorschach.* Seymour B. Sarason. New York: Harper & Brothers, 1954. Pp. 369. \$5.00.
- Personnel management.* Walter Dill Scott, Robert C. Clothier, and William R. Spriegel. New York: McGraw-Hill Book Company, 1954. Pp. 690. \$6.50.

- Personal adjustment in the American culture.* Franklin J. Shaw and Robert S. Ort. New York: Harper & Brothers, 1953. Pp. 388. \$4.00.
- Groups in harmony and tension.* Musafer Sherif and Carolyn W. Sherif. New York: Harper & Brothers, 1953. Pp. 316. \$3.50.
- Man in society.* George Simpson. New York: Doubleday and Company, Inc., 1954. Pp. 90. \$.95.
- Occupational books: an annotated bibliography.* Sarah Splaver. Washington: Biblio Press, 1952. Pp. 135. \$4.00.
- Annual review of psychology.* Volume 5. C. P. Stone, Editor. Stanford, Calif.: Annual Reviews, Inc., 1954. Pp. 455. \$7.00.
- Contemporary theories of learning.* Louis P. Thorpe and Allen M. Schmuller. New York: The Ronald Press Company, 1954. Pp. 450.
- Manual of psychological medicine.* Third Edition. A. F. Tredgold and R. F. Tredgold. Baltimore: The Williams & Wilkins Co., 1953. Pp. 328. \$7.00.
- The juvenile offender.* Clyde B. Vedder. New York: Doubleday and Company, Inc., 1954. Pp. 510. \$6.00.
- The miniature social situation.* W. Edgar Vinacke. Honolulu 14, Hawaii: University of Hawaii, 1954. Pp. 32. \$.65 plus postage.
- Statistical methods in educational and psychological research.* James E. Wert, Charles O. Neidt, and J. Stanley Ahmann. New York: Appleton-Century-Crofts, Inc., 1954. \$5.00.
- Free and unequal: the biological basis of individual liberty.* Roger J. Williams. Austin: University of Texas Press, 1954. Pp. 177. \$3.50.



## Social Status of Industries

Arthur H. Brayfield  
Carroll E. Kennedy, Jr.

*Kansas State College*

and

William E. Kendall

*Chesapeake and Ohio Railway, Cleveland, Ohio*

In 1925, Counts demonstrated that occupations may be arranged in order of social prestige (3). Greatest prestige is usually associated with the professional and "higher" business occupations. Skilled trades, technical, and distributive occupations occupy an intermediate position followed by the semiskilled and unskilled occupations ranked at the bottom of the hierarchy. Research on the social status of occupations has continued and the hierarchical arrangement has been well established. The Counts study was repeated with minor variations 21 years later by Deeg and Paterson who found almost no change in the social status rankings during the intervening years (4).

Does a social status hierarchy exist among industries? It occurred to the writers that an investigation of this question might be interesting and significant. A review of the literature revealed no such studies. In this paper we report an exploratory attempt to ascertain: (1) whether or not an industrial hierarchy exists; and (2) the possible influence of occupational status stereotypes upon the identification of such a hierarchy.

### Method

The base method for this investigation was a ranking procedure similar to that of the studies of occupational prestige hierarchies and closely patterned after Baudler and Paterson (1). An alphabetical list of 29 industries was presented to 68 men and 52 women members of the same class in Gen-

eral Psychology with instructions to "rank according to what you think their social standing is in your community or state." At least one industry from each of the 9 major divisions in the Standard Industrial Classification Manual (5) was included. Competitive industries were included in a few instances as, for example, bus companies, air transport, railroads, and trucking companies. The respondents were predominantly college freshmen and sophomores representing 26 different curriculums. The median rank and its quartile deviation were computed for each industry and the industries were then placed in rank order according to their median values. The rank order correlation ( $\rho$ ) between men and women rankings was computed.

A subsidiary problem was to attempt to discover whether or not respondents were influenced by the social status of a particular occupational level stereotype which might be associated with any given industry. The method employed was to vary the instructions to four additional groups of respondents who ranked the same list of industries. A total of 48 men and 76 women from classes in General, Educational, and Social Psychology responded to instructions to rank the 29 industries "according to what you think the social standing of an executive in each of the industries is in your community or state."

An additional 48 men and 66 women from General and Educational Psychology ranked the industries under instructions to "rank according to what you think the social stand-

Table 1

Rank Order of 29 Industries Based on Median Social Status Rankings by 68 Men and 52 Women College Students\*

Industry	Men			Women		
	Median Rank Order	Median Ranking	Quartile Deviation	Median Rank Order	Median Ranking	Quartile Deviation
Medical services	1	2.1	1.9	1	2.1	1.5
Banks	2	2.7	1.3	2	2.6	1.5
Education	3	4.8	2.7	3	3.6	1.4
Federal government	4	5.4	3.9	4	4.0	2.8
Farming	5	8.5	8.7	5	7.1	6.4
Local government	6	10.8	6.7	6	8.5	5.3
Aircraft manufacturing	7	11.5	4.7	16	14.8	5.3
Broadcasting companies	8.5	12.5	6.3	7	9.5	3.7
Real estate companies	8.5	12.5	5.5	8	10.8	4.9
Air transport companies	10.5	13.2	3.8	11.5	13.8	5.3
Electric light companies	10.5	13.2	4.8	11.5	13.8	3.9
Automobile manufacturing companies	12	13.5	5.7	20	17.5	5.6
General building construction	13.5	14.0	4.8	13	14.2	6.6
Telephone companies	13.5	14.0	5.0	14	14.3	3.1
Chemical manufacturing companies	15	14.3	4.9	15	14.5	6.6
Machinery manufacturing companies	16	14.5	5.4	21	18.5	5.1
Food manufacturing companies	17	15.3	4.9	18	16.5	3.9
Publishing companies	18	15.8	6.2	9	12.5	5.0
Motion picture companies	19	16.3	7.9	10	13.5	5.8
Railroads	20	16.7	6.8	18	16.5	5.9
Retail drug companies	21	18.5	4.6	18	16.5	5.1
Furniture manufacturing companies	22	18.7	4.2	24	20.4	4.1
Wholesale drug companies	23	19.3	3.6	23	19.8	3.8
Hotels	24	21.0	4.9	25	21.2	5.0
Oil drilling companies	25	21.5	6.8	26	22.0	5.8
Bus companies	26	22.0	4.8	22	19.5	4.9
Trucking companies	27	23.7	3.9	27	26.0	2.4
Laundries	28	27.0	2.2	28	27.4	2.0
Coal mining companies	29	27.0	2.7	29	28.2	1.6

\* Median rankings and quartile deviations reported to one decimal place only although median rank orders and computation of rho's were based on medians carried to two decimal places.

ing of a *laborer* in each of the industries is in your community or state."

The results for the latter four groups were treated statistically as for the two base method groups and intercorrelations among the three methods were computed by sex.

### Results

The results of the rankings by the base method groups are shown in Table 1. The median rankings of industries distribute themselves over a wide range (from 2 to 27) whereas chance responses would yield a

clustering around the median value of 14.5. It is evident from inspection of the quartile deviations that there is much greater agreement on the industries ranked extremely high and low than on those ranked in the middle of the distribution.

The correlational results by sex for the three ranking methods are summarized in Table 2. The correlations are all significant beyond the 1% level. The influence of occupational stereotype is small since the correlations are of substantial magnitude.

Men and women agreed markedly in their

status rankings irrespective of method. For the base method the rho was .90, for "Executive," .90, and for "Laborer," .93.

The existence of an industrial status hierarchy seems to be well established by the results from the administration of the three lists. The assignment of ranks is obviously not a chance phenomenon and is relatively uninfluenced by sex.

Table 2

Intercorrelations (rho) between Three Methods of Ranking 29 Industries on Social Status, by Sex

Note: In upper right-hand half of the table the intercorrelations for the rankings by men are given and in the lower left-hand half of the table the intercorrelations for the rankings by women are given.

	Method		
	Base	"Executive"	"Laborer"
Base	—	.89	.89
"Executive"	.78	—	.81
"Laborer"	.92	.84	—

The determinants of such a prestige hierarchy are obscure. For example, the high rank accorded to farming by all groups in this study may reflect a geographical factor. We attempted to ascertain the influence of a possible occupational level stereotype upon the rankings but found little influence within the limitations of the method used. Since the operation of at least a white collar-blue collar stereotype seems probable from an inspection of the rankings a more intensive study of this factor might well be undertaken.

There were interesting differences within a broad industrial classification. For example, furniture manufacturing did not rank higher than 20th on any of the lists while aircraft manufacturing ranked below 8th on only one of the six rankings. In the field of transportation, bus companies and trucking companies consistently ranked well toward the bottom while other forms of transportation enjoyed a considerably higher status. On the

other hand, there was no reliable differentiation between electric light and telephone companies selected as representative of utilities.

The findings of this study should be of interest to several groups. A few industries have demonstrated their concern for public opinion by conducting confidential surveys of the public's attitude toward them. The so-called institutional advertising campaigns are further evidence of this concern. Personnel workers are aware of the influence of public opinion on their recruiting programs (2, p. 88). Further, the prestige associated with an industry may be a factor in job satisfaction.

Vocational counselors should be alert to the possible influence of the industrial status hierarchy on the vocational plans of their counselees. The methodology employed is potentially useful to placement officers in schools, colleges, and public and private employment offices.

### Summary

The existence of a prestige hierarchy among industries was established through the use of a ranking method employed with college undergraduates representative of a variety of curriculums. The influence of occupational level stereotypes was studied and found to be negligible for the populations studied and the method used.

Received August 27, 1953.

### References

1. Baudler, Lucille and Paterson, D. G. Social status of women's occupations. *Occupations*, 1948, 26, 421-424.
2. Bellows, R. M. *Psychology of personnel in business and industry*. New York: Prentice Hall, 1949.
3. Counts, G. S. The social status of occupations: a problem in vocational guidance. *Sch. Rev.*, 1925, 33, 16-27.
4. Deeg, Maethel E. and Paterson, D. G. Changes in social status of occupations. *Occupations*, 1947, 25, 205-208.
5. *Standard industrial classification manual. Manufacturing industries* (Vol. I) and *Non-manufacturing industries* (Vol. II). Washington: U. S. Government Printing Office, 1942.



## Manager-Employee "Understanding" in the Retail Grocery and Meat Market

Pietro V. Marchetti

*University of Illinois*

We have chosen, somewhat arbitrarily, the term "understanding" as a label for the trait or ability of interest to us in this study. This term has been used previously by others as we are using it. Still other investigators have used the words empathy, social psychological empathy, and social perception. The ability we are interested in is that of being able to place one's self in the position of another. We have taken as an indicant of it simply the accuracy with which one person is able to predict the responses that another will make to some given stimulus situation. The now very popular technique, and the one we have employed, is to have one person predict the responses of another to some paper-and-pencil device. Rating scales and personality questionnaires are instruments that may be so employed.

In much common sense speculation about determiners of effective interpersonal relations we find this ability to take another's role (or to take another's point of view) suggested as an important one. Speculations of the social scientist, too, suggest such an ability as an important one in interpersonal relations. The kind of interpersonal relation of interest to us in the present study is that between the formal face-to-face leader and his followers (or subordinates) in the work (or job) situation. Stogdill in his 1948 survey of leadership studies in which some attempt had been made to determine the traits or characteristics of leaders, concluded that this approach to the study of leadership was an inadequate one. "[Leadership] appears rather to be a working relationship among members of a group, in which the leader acquires status through active participation and demonstration of his capacity for carrying cooperative tasks through to completion. Significant aspects of this capacity appear to be intelligence, alertness to the needs and motives of others, and insight into situations, . . ." (16, p. 66).

Gibb has prepared a survey of those leadership studies emphasizing the interactional relationship between the leader's traits and the characteristics of the particular situation in which he functions. Gibb writes, "The function of the leader is to embody and give expression to the needs and wishes of the group and to contribute positively to the satisfaction of these needs" (6, p. 20 f.). Roethlisberger (13) and Barnard (1), among many others who have written in the area of industrial leadership, point to such an ability as an important one in effective leadership. We might note parenthetically that Barnard, a professional industrial manager, wrote in 1940, "Leadership appears to be a function of at least three complex variables—the individual, the group followers, the conditions" (1, p. 16). From his observations in the industrial enterprise he arrived at a statement about leadership quite in accord with the psychologist's interactional theories of leadership which have replaced earlier trait theories.

One psychological study we would note briefly, which served as a major impetus to our own work, is that of Chowdhry (4). She has suggested *situational-traits*—traits common to leadership and yet a function of the situation. She found, in general, that the sociometric leader in the groups she studied (primarily college student groups) could make more accurate judgments or estimates of group opinion than could the non-leaders, defined sociometrically.

The studies of Meyer (12) and of Cantor (3) are two studies, from the pertinent literature, very closely related to our own. Meyer studied 200 first-line supervisors in a large utility company. He asked the supervisors to predict the behavior of other persons, who had been described briefly for them, in certain interpersonal situations. There was evidence that the better supervisors regarded others as individuals with motives, feelings

and goals of their own. The poorer leader was more likely to perceive others in relation to his own motives or goals. Cantor did an experimental study of a human relations program in the Farm Bureau Insurance Companies in Ohio. One of his findings was that following the supervisory training conferences the supervisors showed gains in scores on a test of the ability to estimate group opinion. Lastly, in this sampling of earlier studies which relate to the suggested factor of the leader's understanding of the followers there are two University of Michigan Survey Research Center studies of productivity, supervision and morale. One of these was carried out in a large life insurance firm (7) and the other with gangs of men who maintain sections of railroad right of way (8). In each of these studies there was evidence that employee groups of higher productivity were under supervisors or foremen who were more employee centered. There was evidence of their taking more interest in their employees than did the leaders of lower productivity units. There was some evidence that they considered the possible needs and motives of their employees in their interpretations of the behavior of employees. This last finding may be compared with the results of a study of Mass (11). He reports that the leaders of youth groups sponsored by community agencies, after taking courses intended to make for more effective leadership of youths, made more of what he calls causal reactions rather than judgmental reactions to the behavior of the youths. He gives as an example of a C-reaction, "Joe is smoking a pipe perhaps because he is the smallest boy in the group or perhaps to rebel against paternal sanctions against smoking"; and as a J-reaction, "Joe is a bad boy," or "Joe's only fault is smoking."

Granting as a factor in effective leadership, the leader's ability to understand the group members, a number of questions about this ability quickly arise. One is suggested by Sol Levine's (9) discussion of leadership. There may be a curvilinear relationship between understanding and leadership effectiveness. With too limited understanding we may have Levine's formalistic leader who is not very successful in motivating the group members and eliciting from them a genuine contribution of

their efforts to the group's tasks. The opposite extreme of understanding may make for Levine's anarchic leader acutely sensitive to the feelings of the group members but incapacitated by his lack of ability to abstract, to see beyond the concrete. Apart from the question of optimal amount of understanding there is also the question about the kind of understanding. That is to say, there are many different things that one might know about another person—many different aspects of another's personality that one might understand or know about. It is reasonable to assume that the various possible kinds of understanding that one might have of another are not equally important in the leader-follower relationship in the job situation. Pertinent to this question is work of Luszki (10) on empathic ability and social perception. She presents evidence of some independence between the ability of one person, A, to predict the responses of another, B, to a stimulus situation not involving A, and A's ability to predict the responses of B to a stimulus situation which does involve A. She speaks of *detached* as compared with *participant* observer skill. We shall borrow these terms and use them analogously as adjectives for understanding. A third question we ask is that of differences among work situations in terms of the degree to which effective leadership in the situation is determined by or associated with the leader's understanding of the group members. Where the group tasks are such that individuals function more as automatons the leader's understanding of the group members may be of less importance in leadership effectiveness, particularly so if we take some aspect of group productivity as a criterion of leadership effectiveness. There is a fourth and final question we would raise at this point. This has to do with the relation of "apparent" to "real" understanding. The latter is the sort of understanding with which we are concerned in the present study. This is related to the amount and kind of knowledge that one person has of another which makes it possible for him to make an accurate prediction of how the other person will respond to a given situation. The person A may have such understanding of B. B, however, may not have such understanding of C



and yet he may appear to. A, with the knowledge or understanding that he has of B can select, with a minimum of trial and error, the appropriate stimulus situation with which to confront B in order to elicit a given kind of response from B. B on the other hand does not have such understanding of C. Nevertheless, he is able to elicit, from C, a desired response. B's success in this may be primarily a matter of trial and error. He may be able to confront C with one stimulus situation after another noting very quickly any immediate cues that C may be giving him, on the basis of which B can predict what C's more complete response to the stimulus situation would be. On this basis B can decide if it will be necessary to present C with still some other stimulus situation in order to elicit the desired response or not. Another recognition of this problem is to be found in a discussion of leadership by Smith (15).

The task we have set for ourselves is that of making a more frontal attack upon the problem of leader-follower (and follower-leader) understanding in the job situation. We hope to determine in various kinds of job situations those variables in the job situation which are correlates of the understanding (or rather, understandings) between employee and immediate supraordinate. We are principally concerned, of course, with the degree to which such understandings may correlate with job satisfaction, that is, attitudes of rank-and-file toward various aspects of the job situation; with various criteria of leadership effectiveness; and finally with group effectiveness (or productivity) and group efficiency. We would expect such understanding to be more highly correlated with group efficiency than with group effectiveness. As we have noted earlier, effectiveness or productivity in the job situation is today, many times, more a function of technological factors rather than of the kind of interpersonal relationship between worker and immediate supraordinate. The obvious difficulty in attempting to demonstrate the correlation between understanding and group efficiency is that of developing an adequate indicant of efficiency of the group—an indicant that would reflect the psychological costs, to the

individual worker, of the work accomplished by him. We are thinking here of Ryan's discussion of cost of work to the individual in his *Work and effort* (14). We have also in mind Barnard's discussion of the relation of efficiency of the group to the individual efficiencies of its members, in his *The functions of the executive* (2). It is hoped that ultimately such studies might contribute to more effective training as well as selection of persons to serve in supervisory capacities in various work situations. The results of such studies may also contribute to more effective matching of employee and supervisor. We may eventually be able to consider in the placement of personnel an additional variable and that would be the degree to which the employee might be expected to be enigmatic to a particular supervisor (or the supervisor enigmatic to the employee). The objective, of course, would be to so match employee and supervisor that there might be adequate understanding one of the other.

### Procedure

*Subjects.* In the present study the subjects were the rank-and-file employees and managers of ten grocery retail units and two retail meat market units of a midwestern grocery "chain." One of the grocery units is excluded from the data analysis. It was the first unit in which we collected our data. It soon became apparent in this unit that the questionnaires we had selected for our study were too lengthy. After making modifications of our questionnaires the study was continued in the remaining units. The data collected in the first unit, of course, remain incomparable to those gathered in the other units. The number of employees in the eleven units upon which this report is based, varies from three to 42. More specifically, Unit A had 17 employees; B, 21; C, 3; D, 4; E, 6; F, 4; G, 8; H, 12; I, 42; J, 8; and K, 14. The Units J and K are meat markets.

*Measures of Understanding.* As indicated earlier, the measures, generally, are statements of the accuracy of one person's predictions of the responses of another to a questionnaire. A predicts the responses of B to the items of a questionnaire. For each item there are several response categories, in some order, from 1 to 5. If B chooses the response category 4, and A predicts that B's choice will be 2, we shall say that A's error for that item is 2. The direction of the error was ignored in the present study. One can then determine for A, the mean error score; that is, the mean of the errors made on each



item. All of our understanding measures are just such mean error scores.

The four measures of understanding were the manager's detached understanding of the employee (MDU), the manager's participant understanding of the employee (MPU), the employee's detached understanding of the manager (EDU), and the employee's participant understanding of the manager (EPU). For each employee in each unit we determined an MDU, MPU, EDU, and EPU. The MDU is the mean number of errors made by the manager in his predictions of the employee's responses to the items of the *Tear Ballot for Industry*. The MPU is the mean number of errors made by the manager in his predictions of the responses of the employee to a questionnaire we have labeled *Supervisory Practices Questionnaire*. This is simply a shortened form of an instrument developed by Fleishman (5). The employee's responses to the items of this questionnaire indicate how the employee thinks that his manager typically behaves toward his employees. A sample item is, "He criticizes people under him in front of others." The response categories are: 1. Often; 2. Fairly often; 3. Occasionally; 4. Once in a while; and 5. Very seldom. The EDU is the mean number of errors made by the employee in his predictions of the manager's responses to the *Supervisory Practices Questionnaire* with the items so reworded that the manager's responses indicate how he thinks that he, the manager, typically behaves toward his employees. The EPU is the mean number of errors made by the employee in his predictions of his rating by the manager. The manager rated each employee on each of the following seven characteristics: (1) how the employee receives orders and suggestions; (2) customer relations; (3) initiative; (4) acceptance by fellow workers; (5) promotability; (6) personal appearance; and (7) general effectiveness in present job. These characteristics were suggested in descriptions of poor and good employees which were obtained in interviews with two of the managers.

In each of the eleven units we determined, for each of the four understanding measures, a split-half (odd-even) reliability to which we applied the Spearman-Brown formula to obtain an estimate of the reliability of a test doubled in length. For each of our four measures, then, there were eleven estimates of reliability, one obtained in each unit. The median reliability estimates for the measures MDU, MPU, EDU and EPU are, respectively, .78, .82, .79, and .83.

### Results

Our results are given in the form of rank-order coefficients of correlation between each of the four measures of understanding and (a) the manager's ratings of the employees; (b) the evaluation of the manager by the em-

ployees on the *Supervisory Practices Questionnaire*; (c) the job satisfaction of the employees as measured by the *Tear Ballot*; and (d) the efficiency of the unit as evaluated subjectively by a member of management supraordinate to the unit managers. Each correlation coefficient is based upon eleven cases; the eleven units ranked in terms of the mean MDU of the unit, the mean MPU, the mean EDU, and the mean EPU. The units were, of course, also ranked in terms of the mean ratings by the manager of employees in the unit; the mean evaluation of the manager by the employees in the unit; the mean job satisfaction of the employees; and finally the units were placed in a rank order of efficiency by the managers' supraordinate.

There were no well founded hypotheses about the direction of correlation between the understanding measures and the ratings of employees nor about the direction of correlation between these measures and the employees' evaluations of the managers. For this reason the so-called two-sided test of significance of the correlation coefficient is considered appropriate. For eleven cases the rank-order coefficient must be .60 for significance at the five per cent level of confidence and .74 for significance at the one per cent level. We did hypothesize positive correlations between the measures of understanding and employee satisfaction as well as between these measures and the efficiency ratings of the units. These relationships are suggested both by common sense speculation and earlier empirical studies. To test the significance of these correlations we have used the one-sided test of significance. For eleven cases the rank-order coefficient must be .54 for significance at the five per cent level of confidence and .73 for significance at the one per cent level. These results are summarized in standard type, in Table 1.

It quickly becomes apparent from the data in Table 1 that there is no significant correlation between any of the understanding measures and either the ratings of employees or evaluations of managers by employees. Employee job satisfaction, on the other hand, does seem to have some correlation with the manager detached and participant under-

Table 1

Correlations Between Understanding Scores and (1) Employee Ratings by the Manager; (2) Evaluation of the Manager by the Employees; (3) Job Satisfaction of the Employees; and (4) Efficiency of the Retail Unit

## Employee Ratings and

MDU	.10	.52
MPU	.08	.16
EDU	.07	.43
EPU	-.10	-.08

## Job Satisfaction of Employees and

MDU	.56	.55
MPU	.53	.49
EDU	.43	.48
EPU	.13	.12
MDU/SL	.07	-.13
MPU/SL	.72	.74

## Evaluation of Manager and

MDU	.00	-.26
MPU	-.02	-.22
EDU	.37	.20
EPU	.07	.11

## Efficiency of the Retail Unit and

MDU	.20	.23
MPU	.55	.62
EDU	.15	.22
EPU	.52	.96
MDU/SL	.55	.68
MPU/SL	.61	.63

standing of the employees. The correlation between MDU and job satisfaction is significant at the five per cent level (using the one-sided test) and the MPU and job satisfaction correlation very closely approaches significance at the five per cent level. There is but the suggestion of correlation between EDU and job satisfaction.

In each unit we were able to identify one to three people as the one(s) receiving the greater proportion of choices or votes on a sociometric questionnaire. The sociometric criterion question asked of each employee in each unit, was answered by the employee's singling out the one of his fellow employees whom he would most like to have go with him if he were to be transferred to another unit in the same company. In each unit we determined the mean MDU and MPU based not upon all of the employees in the unit, as we did originally, but now based only upon the most frequently chosen persons in the

unit. We may think then of the manager's detached as well as participant understanding of the sociometric leaders in his unit—MDU/SL and MPU/SL, respectively. MDU/SL has no significant correlation with the mean job satisfaction (the mean of all employees' job satisfaction scores, as originally determined). The MPU/SL, however, does correlate significantly with job satisfaction, almost at the one per cent level. In the present study, then, we find that the greater the accuracy of the unit managers in predicting how they are evaluated by those employees in their respective units, who are most frequently chosen on a sociometric questionnaire, the greater the mean job satisfaction of the unit as a whole.

The efficiency ratings of the units by the managers' supraordinate correlate significantly (at the five per cent level) with MPU, MDU/SL, and MPU/SL. Their correlation with EPU closely approaches significance at the five per cent level.

Turning again to Table 1 and noting the italicized coefficients we find that these values (with two or three exceptions) are relatively of the same order of magnitude as the coefficients discussed earlier. The coefficients in italics are based upon the nine grocery units. The two meat markets are excluded. Obviously, with the change in the number of cases the values of correlation coefficients for the two levels of confidence, which we cited above, do not apply here. The major differences between the two sets of coefficients are that the correlations of MDU and EDU with the employee ratings more closely approach statistical significance with the meat markets excluded; and the correlation between EPU and the efficiency ratings of the units becomes appreciably greater. However, these differences do not appear to be such that they suggest that the grocery and meat market units differ significantly in terms of the relationships explored in this study—the relationships between measures of understanding and employee ratings, manager evaluations by employees, employee job satisfaction, and ratings of efficiency of the units. The sample of meat markets numbering but two made it impracticable to test this statistically. It is proposed to explore



this matter further with additional grocery and meat market units, preferably including units from other companies in order that there may be some test of the generality of the relationships that did emerge—or were suggested by the results of the present study.

Our results appear to us to suggest that some further work along the general lines of this study is warranted. We should like to employ the same as well as some different measures of understanding between leader and follower in different work situations. It would seem profitable to use other than a global measure of job satisfaction; that is, measures of different aspects of job satisfaction (or morale) in order to determine how different measures of understanding may relate differently to the various factors of job satisfaction. We should also like to explore the relationships between measures of understanding and productivity as well as efficiency of the group. We have certain reservations about the efficiency ratings of the present study. We do know that the managers' supraordinate who did the ratings felt that he should not be too influenced by differences among the units in terms of net profits. There are factors determining these profits which are beyond the control of the personnel of the unit. A principal one is that at higher levels of management it may be decided to price merchandise differently in different units, in attempts to determine optimum prices. The rater reported as one basis of his judgments, the criticisms, favorable and unfavorable, made by customers to the company about the service and personnel of the different units. The rater also considered the suggestions originating with the personnel of the different units for the improvement of the operation of the units. Another consideration was the physical appearance of the store—its cleanliness and the effectiveness of the displays of merchandise.

### Summary

Certain earlier studies suggesting the present one have been reviewed very briefly. Several measures of understanding between manager and employees in the retail grocery and meat market have been described. The correlations of these measures with the following variables have been reported: (1)

manager's rating of the employees; (2) employees' evaluation of the manager; (3) job satisfaction of the employees; and (4) ratings of efficiency of the units. None of the measures of understanding correlated significantly with either the first or second of the above variables. Certain ones of the understanding measures did correlate significantly with the third and fourth variables.

Received August 6, 1953.

### References

1. Barnard, C. I. The nature of leadership. In Hoslett, S. D. (Ed.), *Human factors in management*. New York: Harper & Bros., 1946.
2. Barnard, C. I. *The functions of the executive*. Cambridge: Harvard Univ. Press, 1950.
3. Cantor, R. R., Jr. *An experimental study of a human relations training program*. Ph.D. thesis, Ohio State Univ., 1949.
4. Chowdhry, K. *Leaders and their ability to evaluate group opinion*. Ph.D. thesis, Univ. of Mich., 1949.
5. Fleishman, E. A. "Leadership Climate" and supervisory behavior. Personnel Research Board, Ohio State Univ., 1951.
6. Gibb, C. A. The research background of an interactional theory of leadership. *Aust. J. Psychol.*, 1950, 2, 19-42.
7. Katz, D., Maccoby, N., and Morse, N. C. *Productivity, supervision and morale in an office situation*. Ann Arbor, Mich.: Survey Res. Center, Inst. Social Res., Univ. of Mich., 1950.
8. Katz, D., Maccoby, N., Gurin, G., and Floor, L. G. *Productivity, supervision and morale among railroad workers*. Ann Arbor, Mich.: Survey Research Center, Inst. Social Research, Univ. of Mich., 1951.
9. Levine, S. An approach to constructive leadership. *J. soc. Issues*, 1949, 5, 46-53.
10. Luski, M. B. *Empathic ability and social perception*. Ph.D. thesis, Univ. of Mich., 1951.
11. Mass, H. S. Personal and group factors in leaders' social perception. *J. abnorm. soc. Psychol.*, 1950, 45, 54-63.
12. Meyer, H. H. *An investigation of certain factors related to quality of work-group leadership*. Ph.D. thesis, Univ. of Mich., 1949.
13. Roethlisberger, F. J. Understanding: A prerequisite of leadership. In McNair, M. P., and Lewis, H. J. (Eds.), *Business and modern society*. Cambridge: Harvard Univ. Press, 1938.
14. Ryan, T. A. *Work and effort*. New York: Ronald Press, 1947.
15. Smith, M. Leadership: The management of social differentials. *J. abnorm. soc. Psychol.*, 1935, 30, 348-358.
16. Stogdill, R. M. Personal factors associated with leadership: A survey of the literature. *J. Psychol.*, 1948, 25, 35-71.



## An Experimental Evaluation of the Sensitivity of the Empathy Test

Arthur I. Siegel

*Institute for Research in Human Relations, Philadelphia, Pa.*

The Empathy Test (1) is now of interest because of the recently reported high correlations (2) of this test with merit rankings of sales-managers' rankings of automobile salesmen ( $r = .71$ ) and with actual sales records of automobile salesmen ( $r = .44$ ). The test correlated (3) as follows with six criteria of success for union business agents: record for settling grievances and disputes,  $r = .64$ ; recruitment of new members,  $r = .60$ ; per cent vote received in union elections,  $r = .38$ ; enforcement of rules and regulations,  $r = .44$ ; leadership rank,  $r = .67$ ; knowledge of supervisory principles,  $r = .55$ . The multiple R with these six criteria was .76.

The authors of The Empathy Test define empathy in the following way: "This unique talent, well known among 'natural' leaders, successful sales managers, and outstanding counselors, is the ability to 'put yourself in the other person's position, establish rapport, and anticipate his reactions, feelings, and behaviors.' This ability is known as *empathy*, except that the past accepted definitions of empathy seem somewhat inadequate since they stress mere identity of feeling and omit the practical element of prediction of the other's behavior . . . *individuals who are superior in empathetic ability are persons who are above average in understanding and anticipating reactions of other people*" (emphasis ours).

The Empathy Test consists of three sections. In the first section the respondent is asked to rank the popularity of 14 musical types (polkas, classicals, waltzes, etc.) with non-office factory workers of the United States. In the second section, the respondent ranks the popularity of 15 magazines with the average American, and in the third section the respondent ranks the annoyance magnitude of 15 experiences (a boisterous person attracting attention, hearing a person chewing gum, seeing a person's nose running,

etc.) to persons aged 25-39. Thus, in all of the sections the respondent is asked to reply not as he would answer, but as the average person would perform the ranking, and from these rankings an empathy score is derived.

Although some low correlations have also been reported by Kerr and his co-workers (1), in view of the high correlations obtained it seemed that some independent experimental evaluation of The Empathy Test was warranted. Assuming the validity of The Empathy Test and assuming that clinical psychologists are higher on empathy than experimental psychologists, then clinical psychologists should score higher on The Empathy Test than experimental psychologists. This assumption for clinical psychologists does not seem to be outside the scope of definition of empathy as given by the authors of The Empathy Test, and seems tenable to the present author.

### Method

Form A of The Empathy Test was distributed by mail to 50 "fellows" of the Division of Experimental Psychology and 50 "fellows" of the Division of Clinical and Abnormal Psychology of the American Psychological Association. The sample was obtained by taking every fifth "fellow" listed in the 1951 A.P.A. Directory in the Division of Experimental Psychology and every tenth "fellow" in the same directory in the Division of Clinical and Abnormal Psychology until a total of 50 names in each division were obtained. In some instances, no clear address was listed and in that case the name appearing directly below the ordered name was used. A total of 36 of the forms were returned by the "experimentalists." Of these, only 34 were completely filled out and one was received after our data were already analyzed. Thus, our total N for experimentalists was 33. A total of 25 out of 26 of the forms re-

Table 1

Means and Sigmas of Clinicians and Experimentalists on Empathy Test

	Mean	Sigma
Clinicians	87.7	14.7
Experimentalists	86.7	18.1

turned by the "clinicians" were usable ( $N = 25$ ). None of the subjects were informed of the purpose of the experiment until after all of the forms used in the comparison had been returned.

### Results

The Empathy Tests were scored and means and standard deviations calculated. These data are presented in Table 1.

The mean Empathy Test score for "experimentalists" was 86.7 while the mean Empathy Test score for "clinicians" was 87.7. The difference between the means is not significant. The mean scores obtained would place both the "clinicians" and the "experimentalists" at the 70th percentile on The Empathy Test's norm for college men.

Since liberal arts female students score lower on The Empathy Test than liberal arts males, and since 14 female "clinicians" were sent the test while only one female "experimentalist" received the questionnaire, the objection may be raised that this sampling differential operated so as to bias the scores of the groups in favor of the "experimentalists." However, if this were the case, the variance of the clinical group should have been greater than the variance of the experimental group. The reverse was true.

All of this might indicate that The Empathy Test either measures something other than empathy, measures empathy plus another variable, or is not a sensitive instrument.

An alternative explanation has been advanced by Kerr, who kindly reviewed an early form of the present paper. Kerr points out the possibility that the better clinicians, possessing a vested interest, may have been more defensive about "going out on a limb" and thus the better clinicians may not have returned the forms. This sampling differential may have acted to lower the empathy scores of the "clinicians." The present author feels that this explanation is unwarranted in view of the fact that neither group was informed of the purpose of the research until after the forms were returned. If the clinicians were unaware of the purpose of the research, there was little reason for the better clinicians to believe that they were "going out on a limb," and thus withhold the returning of their forms.

In fairness to the authors of The Empathy Test, we would like to point out that they have never claimed that it will distinguish between clinical and experimental psychologists. Moreover, the assumption that clinical psychologists are higher on empathy than experimental psychologists was our assumption.

### Summary

The Empathy Test was submitted by mail to a group of experimental and a group of clinical psychologists. Assuming that the "clinicians" are higher on empathy than the "experimentalists," The Empathy Test did not reflect this difference.

Received August 24, 1953.

### References

1. Kerr, W. A. and Speroff, B. J. *Manual for the Empathy Test*. Chicago: Psychometric Affiliates, 1951.
2. Tobolski, F. P. and Kerr, W. A. Predictive value of The Empathy Test in automobile salesmanship. *J. appl. Psychol.*, 1952, 5, 310-311.
3. Van Zelst, R. H. Empathy Test scores of Union Leaders. *J. appl. Psychol.*, 1952, 5, 293-296.

## The Validation of an "Indecision" Score for Prediction of Proficiency of Foremen

J. P. Guilford

*University of Southern California*

The results to be reported briefly here are essentially negative, but perhaps negative results should be reported more often than they are. On the one hand such a report may save another investigator from entering the same blind alley. On the other hand it may give another investigator an idea for doing a similar study in a modified way which will lead to positive results.

The study is also opportunistic, in the sense that it was not planned in advance but was possible as a byproduct of another study. The writer happened to have at his disposal the answer sheets from more than 400 foremen in an eastern industrial plant, these foremen having taken the three personality inventories, STDCR, GAMIN, and Personnel Inventory (of factors O, Ag, and Co).<sup>1</sup> The writer had also been supplied with ratings of general proficiency of the same foremen as judged by their immediate superiors. Unfortunately, details we should like to have concerning the administration of the inventories and the way in which the ratings were obtained are seriously lacking. It can only be said that the inventories were administered after the foremen were employed and that the ratings were on a five-point scale, with efforts made to disperse the frequencies toward a normal distribution. There is no information concerning reliability or validity of the ratings. We can only assume that they have some reliability and validity, for they were predictable from inventory scores.

Another study has dealt with the validation of the 13 inventory scores against the rating criterion.<sup>2</sup> The interest in the present report is directed toward individual differ-

ences in the tendency of the examinees to use the question-mark response to the items. It will be remembered that the alternative responses to the items are "Yes," "?," and "No." Each examinee could be given a score according to the number of "?" responses he gave. It was hypothesized, subject to certain qualifications to be mentioned later, that a large portion of the variance of the "?" score represents a personal trait of indecision. The greater the number of "?" responses an individual gives, the greater his degree of indecisiveness. It was also hypothesized that indecisiveness is an unfavorable trait for foremen and it was consequently predicted that the correlation between this score and the criterion would be significantly negative.

The "?" score comes in the general category of response-set scores that are receiving increasing attention as possible objective measures of personality traits. The meaning of such scores, even when they prove to be highly reliable, must always be questioned. While the first hypothesis about the meaning of the proposed score is that it measures a trait of indecisiveness, there can be other hypotheses, which we will consider.

Indecision can enter into the picture in more than one way. Let us assume first that the examinee is cooperative and attempts to answer each item in the way that most nearly describes himself. He is most likely to waver between responses "Yes" and "No" under two conditions. One is when he does not know himself very well with respect to the question asked. Some "?" responses, of course, represent complete ignorance or inability to give one of the other responses. But when there is partial knowledge, whether the examinee will give the "?" response or one of the others will depend upon his readiness to make a more or less arbitrary choice versus his inclination not to do so. This is the kind of case whose behavior one would

<sup>1</sup> The contributor of the data on which this report is based wishes his organization to remain anonymous. I am nevertheless grateful to him for making the data available.

<sup>2</sup> R. R. Mackie. *Norms and validities of 16 test variables for predicting success of foremen*. A Master's thesis in the University of Southern California Library, 1948.



like to measure by means of an "indecision" score.

Another occasion for wavering is when the examinee knows himself well but is himself near the limen for the item; he is on the borderline that to him separates "Yes" and "No." This kind of indecision, too, we would like to have included in the measurement, since it is probably psychologically identical with the first type mentioned. In this connection we have the problem of equality of opportunity for indecision. Presumably, an examinee who is near the limen for most of the items has much more occasion for wavering than an examinee who is decisively on one side or the other of the trait continuum for most items. If a "?" score were based upon the items that are keyed for one trait only, it can be seen that those who earn moderate inventory-trait scores have more opportunity for wavering than those who earn scores at either extreme. The relation between the "?" score and the inventory-trait score would be curvilinear. Since each inventory is scored for relatively independent factors and the intercorrelations of scores tend to be small, it is very unlikely that an examinee will be at moderate positions on all traits. Hence the opportunities for wavering are somewhat equalized if we obtain a "?" score from all the items in the inventory combined. Some index of opportunity might well be taken into consideration, however, if we want variations in "?" scores to represent traits such as indecision, freed from involvement with patterns of factor scores.

A third occasion for wavering and indecision occurs among those examinees who may have decided to answer the items not as they are but as they think will make a good impression. Here the wavering is with respect to which is the more favorable response, "Yes" or "No." It is likely that without knowledge of the key and with lack of experience in taking inventories, many instances of liminal alternatives arise. Again the kind of indecisiveness in which we are interested would have room for play. The "?" responses given under this condition should also indicate the trait we want to measure.

Three possible meanings of the "?" response have been discussed. All of them, it

has been argued, are potentially contributory to the indecisiveness variance we want to emphasize in the score. Other meanings that do not contribute to this variance include cases in which the examinee does not know the answer to the question and he should therefore legitimately respond with the "?," and cases in which he is at or near the limen for the item and a "?" response represents a correct position for him between "Yes" and "No" on the trait continuum. But, as was pointed out above, there are individual differences in tolerance of an indecisive response, and this fact makes the "?" response contribute to the variance we want. On the other hand, there is a possibility that the significant difference here is in the form of willingness to guess or to gamble versus a caution in this regard. This is not logically an aspect of the indecisiveness variable with which we are concerned.

If the lack-of-self-knowledge component of the "?" score is appreciable, it would add to reliability and also probably to validity against the foreman criterion. The greater the lack of knowledge of self, the poorer should be the chances of success as a foreman or of leaders of other kinds. The willingness-to-guess component should add to reliability but its effect on the "?" score (which is to reduce that score) would tend to detract from validity against the foreman criterion, assuming that good foremen are inclined to be cautious in a situation like this.

### Results

Each of the three inventories was administered as a unit and was given an indecision score as a unit rather than factor by factor. This was partly to assure a larger range of scores and partly to equalize opportunity for wavering, as suggested above. It was of interest, first, to determine whether individual differences in indecision scores are consistent from one inventory to another. The intercorrelations of scores from the three inventories provide estimates of alternate-form reliability.

The frequency distributions of the three indecision scores all approached the Poisson type, with modes at a score of zero. The proportions of zero scores were .41, .48, and

.55. This form of distribution was obtained under the pressure of the instruction for the examinee to avoid the "?" response. It was assumed that the underlying trait continuum, however, was one on which the distribution in the population is normal. Tetrachoric correlations were therefore computed. They were .73, .75, and .88, with an average correlation (Fisher-Z method) of .80. Since these intercorrelations were fairly high the three indecision scores were summed to yield one score for each examinee. The reliability of such a score should be in the region of .90.

The correlation of this combined indecision score with the rating criterion was also found by means of the tetrachoric  $r$ . The sample of 405 foremen was divided into two groups, one having to do with tools and maintenance and the other with production. The scatter plots show no signs of non-linearity. The validity coefficients were +.14 and -.09 for the two groups, respectively. With  $N$ s of 119 and 286, these coefficients are statistically insignificant. They also differ in sign. We may therefore accept the idea that they are random deviations from zero correlation and conclude that there is no support whatever for the original hypothesis.

While there is no evidence of validity of the indecision scores in connection with the performance of these foremen, the level of reliability of the scores is promising of a type of personality measurement that has much stability and may be well worth further study. To be useful for practical purposes,

however, something would need to be done to improve discrimination at the lowest levels where discrimination is now very poor. Had there been differentiation among those scoring zero, we might even find some relationship between scores in that range and the criterion. It would be more reasonable to expect the relationship to appear among scores at the upper levels, however, where none was found. Since the reasoning concerning the contributions to variance in the "?" score indicates several possible traits, a factor analysis of the score is definitely called for as a basis for intelligible future predictions.

### Summary and Conclusions

An "indecision" score was obtained by counting the number of "?" responses to items in the Guilford personality inventories STDCR, GAMIN, and Personnel Inventory. The three scores showed an average intercorrelation of .80, indicating that they measure much the same trait or traits. A combination of these three scores correlated +.14 and -.09 with a rating of proficiency of foremen in an industrial plant, whereas a significant negative correlation had been predicted. While the indecision score indicates something stable about individuals, it needs to be factor analyzed to be understood and test conditions that will assure better discriminations at the lower levels are needed for a score of practical use.

*Received September 14, 1953.*

## An Approach to Isolating Dimensions of Job Success<sup>1, 2</sup>

Louis L. McQuitty, Charles Wrigley, and Eugene L. Gaier

*University of Illinois*

Many currently-employed indices of on-the-job performance do not adequately measure job success, because the original job descriptions do not clearly depict the psychological requirements of the job being studied. As research progresses, it is becoming increasingly obvious that the usual types of job descriptions are neither rigorous nor analytic enough to furnish a sound basis for the development of valid measuring devices. There is a need for new methods by means of which the basic dimensions of job success can be isolated and precisely described.

The present study is the first of a series designed to investigate the possibility of deriving meaningful job requirements by statistical rather than by "rational" analysis. The research plan calls for factor-analyzing descriptions of on-the-job behavior obtained by interviewing the peers and supervisors of selected Air Force Airplane and Engine Mechanics.

This procedure was guided by the following working hypotheses:

1. Peers and supervisors can select representatives of three categories of mechanics, viz., best, average, and poorest.
2. Descriptions of representatives of these three categories will reflect individual differences in psychological variables related to job proficiency.
3. Factor analysis of the descriptions will assist in understanding some of the psychological characteristics related to job proficiency.

4. Ability test items can be prepared which measure these psychological characteristics; individual differences in responses to these items will be related to criteria of job proficiency.

If these hypotheses are to prove fruitful, the following conditions must be met: (a) the descriptions of "best" and "poorest" mechanics should differ significantly; (b) the factors deriving from the descriptions should be meaningful; (c) the descriptive factors should be related to independent criteria of job performance, such as their rated job proficiency by other informants; and (d) the use of the factors as guides in preparing ability items should result in tests which are more highly related to criteria of proficiency than those prepared exclusively by way of the job description approach. The first two conditions and a preliminary investigation of the third form the basis of the present study. More thorough investigation of the last two conditions will follow later, provided of course that the results of the present set of studies justify it.

### The Descriptive Inventory

The present paper reports: (a) the preparation of an inventory, called the Descriptive Inventory, designed to facilitate the description of mechanics by their peers and supervisors; (b) a factor analysis of the results obtained when this inventory was used by supervisors to describe individuals whom they had selected as representative of "best," "average," or "poorest" mechanics; and (c) an analysis of the relations of the items to "best" and "poorest" mechanics.

To obtain items for the inventory, experienced mechanics, most of whom had been in supervisory positions, were asked to select the "best" (or "average" or "poorest") A. & E. mechanic with whom they had worked within the last two

<sup>1</sup> This study was supported in part by the United States Air Force under Contract AF 33(038)-25726, monitored by the Commanding Officer, Human Resources Research Center, Attention: Director of Operations, Lackland Air Force Base, San Antonio, Texas. Permission is granted for reproduction, publication, use and disposal in whole or in part by or for the United States Government.

<sup>2</sup> The authors wish to express appreciation to Charles Baldwin, Charles N. Cherry, and K. Patricia Cross for their assistance in collecting and editing the interview materials, to Walter A. Cleven and Malcolm M. Helper for carrying out the statistical analyses, and to Donald R. Shaw for assistance in the factor interpretation.



years. A description was then sought of the behavior of this mechanic both on and off the job, and the descriptions were subsequently divided into separate descriptive phrases. This plan was followed because it was believed that an inventory constructed in terminology familiar to maintenance personnel would be used with more discrimination by mechanics than would one composed of more academic and technical phrases.

*Subjects and Procedure.* A total of 104 students attending Flight Engineering School at Chanute Air Force Base served as subjects for the initial phase of this study. Each subject was individually questioned in an interview divided into two separate sections: (a) a free-response phase, in which the subject was asked simply to describe a fellow mechanic selected by him to represent one of the three categories of proficiency; and (b) a "structured" phase in which comments were elicited in response to specific questions asked by the interviewer.

*The Descriptive Phrases.* To facilitate drawing of items from the interview protocols, each separate and complete idea (which in most instances could be represented by a single phrase) was demarcated from every other idea. This was done for the structured as well as the free-response portion of the interviews.

These descriptive phrases thus obtained were extracted from the typescripts of the recorded interviews and assembled into a pool, which numbered in all some 15,000 items. From this pool, 264 items were selected on a random basis to serve as the raw material of the inventory. Each of these 264 items was then edited and reviewed independently by three psychologists and five mechanics to the end that: (1) the ideas expressed were always those of the interviewee; (2) each phrase was in the present tense; (3) items on which at least two of the three judges could not agree (in terms of meaning, wording, etc.) were eliminated as ambiguous; and (4) phrases whose meanings were dependent upon context were rewritten to make their meanings clear when used in isolation.

### First Pilot Study

Upon completion of the editing, 235 items remained out of the original sample of 264. These items were assembled into an experimental inventory in which each item required either a "yes" or a "no" answer. The inventory was administered to Air Force supervisors in order to: (a) obtain comments as to the meaningfulness and adequacy of coverage of the phrases; and (b) to secure a preliminary indication of the predictive utility of the phrases. To fulfill this latter aim, chi square values were computed for each item in order to determine whether or not it dis-

criminated significantly between mechanics selected as "best" and "poorest."

After these data had been obtained, the items were considered one at a time with respect to: (a) the mechanics' comments; (b) the magnitude of the chi square values; and (c) the proportion of subjects answering each answer alternative. Three judges decided whether to retain, amend, or reject items. Items with a 90% or more response for either answer alternative were rewritten to lessen this percentage whenever this appeared possible; otherwise they were rejected. Items which supervisors reported to require information that they did not have were rejected, and those regarded as difficult to understand were amended.

In all, 35 items were eliminated in this phase of the study. The 200 remaining phrases were assembled in random order into a check list designated as the Descriptive Inventory. Although the entire inventory was used in the collection of data, only the first 120 items were analyzed in this study, for reasons stated later. Examples of the items are listed in Tables 1-2.

### Use of the Descriptive Inventory

Our next immediate purposes were: (a) to isolate relatively independent clusters of interrelated descriptions; (b) to interpret these psychologically; and (c) to make a preliminary investigation of their relation to job proficiency.

*Subjects.* The Descriptive Inventory was administered to 428 Flight Engineering students at Chanute Air Force Base. Each subject had completed a course in Airplane and Engine Mechanics and had had at least six months of supervisory experience. In length of line maintenance experience the subjects ranged from six months to more than 21 years, with a median of four years.

*Administration.* The inventory was administered in small group sessions (12 to 25 men) of about 30 minutes in length. The instructions printed on the face sheet of the booklets were read to the subjects before they began work on the inventory. All of the respondents were given ample time to complete the items.

*Method of Analysis.* In order to reduce computational labor, the Descriptive Inventory was divided into two parts for factor analysis by the shortened square root method developed by Wrigley and McQuitty<sup>3</sup> (a modification of Thurstone's diagonal method). In the present paper, results are reported for the first 120 items only. (Since the 200 items in the Inventory

<sup>3</sup> Wrigley, Charles and McQuitty, Louis L. *The Square Root Method of Factor Analysis: A Reexamination and a Shortened Procedure* (Manuscript).

were arranged in random order, there should be no significant difference in the type of items appearing in the two parts.) Phi coefficients were used to measure correlations; these and the factor loadings were calculated on IBM equipment, using punched card methods developed by Helper.<sup>4</sup>

In the square root factor analysis, a "pivot variable" was selected and factor loadings were calculated to reduce the correlations or residual correlations for that variable to zero. The pivot variables were selected with the aim of enhancing the likelihood of getting predominantly positive factor loadings, and of obtaining factors which are conceptually clear. The pivot variable was always the one with the highest absolute column sum. The same procedure is then repeated with another "pivot variable." The method results in orthogonal factors, with all factor axes at right angles to one another.

In addition to the factor analysis, a pilot validity analysis was also completed. Using only the 204 inventories which described "best" and "poorest" mechanics, phi coefficients were computed between each item and the best-poorest dichotomy. These phi coefficients are here called criterion correlations.

A total of 18 factors were extracted from the 120 variable matrix. By this time, the point of decreasing returns appeared to have been reached, as shown by the drop in relative proportions of variance accounted for by the 15th, 17th and 18th factors; moreover, the factors became less obvious in psychological meaning.

In order to insure that no major group factors had been omitted, a list was prepared of all the items not appearing within the ten highest loadings of any one of the first 18 factors. Further pivots were drawn from this reduced list. This procedure was designed to guarantee that some axes passed through that portion of the hyperplane which had not previously been traversed. The value of the procedure in the present study was demonstrated by the fact that the next factor—the 19th in order of extraction—proved to be the sixth in the order of variance. The next four were less encouraging. Consequently, the factoring was discontinued at this point. This made a total of 23 factors extracted.

<sup>4</sup> Helper, M. M. Punched-card procedures for square root factor analysis (Manuscript).

### Item Validities

It will be of interest first to consider the phi coefficients between the individual items and the "best-poorest" dichotomy for the 204 mechanics classified in this fashion. These criterion correlations range in magnitude from .87 to .01 with a mean of .48. Although these results show some very substantial relationships between the descriptive items and the "best-poorest" criterion, they cannot, of

Table 1

Descriptive Inventory Criterion Correlations for Items with High Predictive Value ( $\phi > .70$ )

Item No.	Phi Coefficient	Item
Items characteristic of good mechanics		
68	.868	He makes sure he does a good job
11	.835	When he does a job you know it will be done right
44	.812	He deserves a promotion
97	.800	If you leave him to do a job, you can always be sure he will get the job done
107	.782	He is good at working on the plane
105	.780	He can show you how to do the job right
34	.772	He is a good man on any job
91	.772	He knows his stuff
119	.768	He seems to take pride in his work
47	.762	You don't have to worry about telling him what to do all the time
72	.759	He tries to find better ways of doing things
106	.739	His ambition will pay off
118	.721	He gives good cooperation
49	.712	He will straighten a guy out and explain things to him
6	.703	If he were a crew chief, he would work right along with his men
Items characteristic of bad mechanics		
12	-.778	You wouldn't feel safe unless you checked behind him
27	-.753	Most guys with that much experience know a lot more than he does
10	-.750	He works in a sloppy way
38	-.748	He isn't a very careful worker
41	-.744	He achieves his aim in the wrong way
69	-.741	He is kind of slipshod in his ways
30	-.714	He doesn't have any sense of responsibility



course, be accepted as final evidence that the descriptions are related to proficiency on the job. They indicate, however, the features regarded by the supervisors as significant.

Items with highest criterion correlations (see Table 1) are mostly generalized descriptions of behavior on the job, e.g., "He makes sure he does a good job"; "When he does a job, you know it will be done right." These characterizations, however, give little detailed information as to the psychological components of job success. The advantage in carrying out a factor analysis of the items is that more analytic dimensions are thus developed.

The items with the lowest criterion correlations (see Table 2) are less job centered in their orientation, and deal with such traits as drinking habits, social demeanor, truthfulness, appearance, etc.

The items with high criterion correlations agree for the greater part with the items with high loadings on the first factor. This is particularly evident if the highest 20 items in each instance are considered. Sixteen items are common to both the lists of the 20 highest factor loadings in the first factor and the 20 highest (absolute) correlations with the criterion. The main differences are: (a) the order of appearance of the items is somewhat changed in the two lists; (b) the highest loadings on the first factor stress the elements of cooperation and dependability, but the corresponding criterion-correlation items are even more generalized, saying little more than that the mechanic does a good job.

The advantages in making the factor analysis are thus clearly seen. If the study had been restricted to criterion correlations, there would have been no concise account of different psychological components related to the pilot criterion. The function of the factor analysis is to aid in identifying some constituents which are involved in the descriptions of mechanics.

### Interpretation of Factors

As computational methods improve and the analysis of more variables for larger numbers of subjects becomes practicable, factor analysts will probably become accustomed to dealing with smaller loadings. In this study,

Table 2

Descriptive Inventory Criterion Correlations for Items with Low Predictive Value ( $\phi < .20$ )

Item No.	Phi Coefficient	Item
54	.014	He is of just average appearance
55	.026	If there is something he likes to do, he does it faster than anyone else would
61	.179	He doesn't lose his temper
37	.192	He would give the shirt off his back
24	.195	He associates with fellows like himself
117	.199	He is of above average appearance
32	-.010	He is quite young for his rank
88	-.028	He doesn't drink
26	-.058	He likes to drink on his off-duty hours
62	-.072	Sometimes he gets "T'd off"
29	-.077	His basic training was rather short
58	-.088	We have had a couple of "run-ins"
92	-.148	He appears very rude to a stranger
96	-.150	He hasn't had too much time in Service
108	-.193	He doesn't care to mix with other people

loadings as low as .10 usually appear to be quite meaningful, and not at all inconsistent with the general interpretation of the factor. The twelve factors accounting for the most variance are reported here. The sums of squares of loadings for each of the factors are presented in Table 3.<sup>5</sup>

Tables 4 through 15,<sup>6</sup> one for each of the twelve factors, report (a) the items with the ten highest loadings for each factor, the positive loadings first, followed by the negative ones; (b) the interpretation of each factor; (c) the phi coefficient between the subjects

<sup>5</sup> In the use of the square root method, the larger factors tend to, but do not necessarily, appear before the smaller. In presenting results here, the factors have been rearranged in order of contribution to variance, and to conserve space, smaller factors have not been reported. The factor loadings are on file for all 23 factors at the Training Research Laboratory, University of Illinois.

<sup>6</sup> Tables 4-15, and 16 and 17 have been deposited with the American Documentation Institute. Order Document No. 4248 from the ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress, Washington 25, D. C., remitting in advance \$2.25 for 35 mm. microfilm or \$5.00 for 6 x 8 in. photocopies. Make checks payable to Chief, Photoduplication Service, Library of Congress.



Table 3  
Square Root Factor Analysis: Sums of Squares of Factor Loadings

Factor in Order of Size	Factor in Order of Extraction	Sum of Squares of Factor Loadings	Factor in Order of Size	Factor in Order of Extraction	Sum of Squares of Factor Loadings
1	1	24.7028	14	22	1.5272*
2	2	2.8732	15	8	1.4942
3	4	2.3592	16	20	1.4780*
4	12	2.3522	17	3	1.4096
5	11	2.0080	18	9	1.3614
6	19	1.9102*	19	16	1.2952
7	6	1.8269	20	21	1.2469*
8	10	1.7407	21	17	1.1953
9	5	1.6646	22	18	1.1584
10	14	1.6644	23	15	1.0812
11	7	1.6302	Total sum of squares		61.1084
12	13	1.5756	Contribution to variance		50.92%
13	23	1.5530*			

\* Pivots for these factors were selected from a reduced list of variables, viz., those which had hitherto not carried very high loadings on any factor.

selected as "best" or "poorest" and the item response; and (d) the code number of the item in the inventory. Loadings as low as .10 were accepted here because of the large number of both variables and subjects ( $N = 428$ ).

Table 16 lists all 120 items of the Descriptive Inventory. Table 17 gives the 12 factor loadings for each of the 120 items. Tables 16 and 17 are also deposited in ADI. (See footnote 6.)

#### Relation of Factors to Criterion

The problem remains as to whether all factors described by the mechanics are related to the criterion. Results may be summarized by presenting the average criterion correlation for the 10 items with highest loadings in each factor. Those factors which appear to be measuring somewhat the same area of behavior have been grouped together.

These results appear quite clearcut. The drive and initiative shown by the mechanic, on the one hand, and his practical efficiency, on the other, are most closely related to the pilot criterion. The seven factors which are grouped under these two headings have the highest average for criterion correlations. The factors dealing with social manner of the mechanic, his interest in aircraft and in the Air

Factor No.	Factor Title	Average Criterion Correlation
<i>Aspects of drive and initiative.</i>		
1.	Sense of responsibility	.74
3.	Willingness for work	.50
4.	Laziness	.59
8.	Industriousness	.58
<i>Aspects of practical efficiency.</i>		
6.	Failure to use knowledge effectively	.71
14.	Practical workmanship	.66
16.	Lack of craftsmanship	.70
<i>Aspects of knowledge and intellectual powers.</i>		
7.	Teaching capacity	.41
9.	Memory	.55
15.	Intellectual capacity	.48
21.	Job knowledge	.47
<i>Aspects of social manner.</i>		
12.	Social acceptability	.54
17.	Personal pleasantness	.35
19.	Anti-sociability	.35
<i>Aspects of interest and morale.</i>		
2.	Interest in aircraft maintenance	.41
13.	Lack of morale	.38
<i>Aspects of character.</i>		
5.	Weakness of character	.38
10.	Self-control	.23
20.	Lack of self-control	.10
<i>Other factors.</i>		
11.	Inexperience	.31
18.	Tendency to mediocrity	.05

Force, and his intellectual powers are less highly related to the criterion. His drinking and money habits, and his ability to control his temper have little or no relation here.

The factors which account for more variance tend to be more highly related to the pilot criterion, as shown by the fact that the rank-order correlation between factor variance and mean criterion correlation for the factors, using all 23 factors, is .51.

### Discussion

*Interest and Motivation in the Supervisors' Accounts.* In discussing these results, we must bear in mind that these were the ratings made by supervisors, and their judgments may reflect their own conceptions rather than actual job performance. In terms of their judgments, interest and motivation appear as the principal factors in the descriptions of "best," "poorest," and "average" Aircraft and Engine Mechanics. Primarily, the mechanic is described as to: (a) whether he is cooperative and can be depended upon to get the job done; and (b) whether he likes being a mechanic and working on aircraft. The mechanic's behavior is reported more frequently than what he knows. Little is said about his technical knowledge; and "poorest" mechanics are frequently described as disinterested or lazy rather than stupid or inadequately trained. The overall picture of the good mechanic is one of being responsible and being willing to work and learn a few simple things, rather than of any extensive knowledge of the principles of mechanics.

*The Place of Mechanical Information.* All Aircraft and Engine Mechanics are supposed to possess at least a minimal amount of technical knowledge. Presumably some mechanical learning is necessary if a man is to be able to service an aircraft, but this study has revealed neither the nature nor the amount of this basic information that is required of the successful mechanic. Supervisors make little reference to lack of this, even in their descriptions of bad mechanics. Hence, we may assume that: (a) the amount needed is less than has generally been considered to be the case; (b) the Air Force is highly successful in giving to all recruits who pass through technical training school the groundwork of knowledge which is prerequisite to satisfactory job performance; or (c) the supervisors

neglected in their descriptions a significant characteristic in which mechanics differ. In any case, the restriction of this study to men already trained minimized the importance of mechanical knowledge, and consequently emphasized differences in motivation and personality. Even if differences of interest and willingness do not give the whole story, this factor analysis has made abundantly clear that they are, at least, the primary variables in the descriptions by supervisors of mechanics whom they selected to represent different levels of efficiency. In other words, the results support the hypothesis that getting good Aircraft and Engine Mechanics is not entirely a problem of accumulating knowledge; it is, at least in part, a matter of motivation, interest, and morale.

### Summary

Before additional tests designed to predict success of mechanics are written, specific hypotheses are needed outlining the dimensions which enter into job proficiency. The present study attempted to isolate some of these hypotheses by: (a) obtaining descriptions from supervisors, in their own words, of Airplane and Engine Mechanics (selected to vary in proficiency); and by (b) factor-analyzing a compendium of these descriptions. A square root factor analysis of 120 of these descriptions resulted in the following hypotheses, for further study.

1. There are a large number of rather independent dimensions of behavior related to job proficiency. Of the 23 factors extracted, practically all of these were found to be related to differences in mechanics selected as representative of "best" and "poorest" job performers by the supervisors who described them.

2. The six most clearly defined of the 23 dimensions were asserted to be: (a) sense of responsibility; (b) interest in aircraft maintenance; (c) willingness for work; (d) laziness and lack of initiative; (e) weakness of character; and (f) failure to use knowledge effectively.

It was concluded that supervisors describe trained mechanics who are selected by them to vary in proficiency much more in terms of interest and motivation than in terms of the amount of job knowledge possessed.

Received August 6, 1953.

# The Analysis of an Experimental Job Evaluation System as Applied to Enlisted Naval Jobs<sup>1</sup>

E. J. McCormick

*Occupational Research Center, Purdue University*

and

Willard E. North

*6563 Research and Development Group, Chanute Air Force Base, Illinois*

There has been a growing recognition within the military services of the potential usefulness of job evaluation for various purposes. Military job evaluation, for example, can contribute to the more adequate differential qualitative allocation of personnel among the services, to the development of career programs, and to the equalization across the services of grades and ranks for comparable types of responsibilities and duties.

The present investigation is a pilot study relating to the job evaluation of enlisted jobs within the United States Navy. Specific purposes of the study were those of identifying the factors which contribute to differences in job values, and of determining the relative importance of each such factor.

<sup>1</sup>This article is based on a study carried out by the Occupational Research Center, Purdue University, under the provisions of a research contract between the Office of Naval Research and the Purdue Research Foundation (Contract No. N7onr-39410). The views expressed herein are those of the authors and do not necessarily represent the views of the Navy Department.

The authors wish particularly to express their appreciation to Mr. D. G. Price, Chief, Billet and Qualifications Research Branch, Personnel Analysis Division, Bureau of Navy Personnel, for his cordial cooperation in making arrangements for many phases of this investigation.

<sup>2</sup>The term "job" does not have a specific connotation in the Navy as it does in industry and business, but is used in this article for purposes of convenience of terminology. Because of the nature of shipboard operations, an enlisted man must perform several different sets of duties at different times under different conditions that are involved in operating and fighting a ship. Thus, an enlisted man usually has certain "routine" duties that he performs (these are the regular duties of the individual in tasks that are related to his Navy rating); he also has certain "Watch, Station, and Quarter Bill" assignments that he performs under specific shipboard conditions, as for example during emergencies or during specified operations. The study reported in this article was based largely on what might be thought of as the "routine" duties of enlisted personnel.

## Experimental Procedures

The experimental procedures basically involved the identification, for one representative sample of enlisted jobs, of the factors (and of their statistically determined weights) which gave the optimum degree of relationship with criterion values, and the cross validation of the results with a second representative sample of jobs. It was hypothesized that if a particular collection of factors, with their appropriate weights, would predict job values with two independent samples, a job evaluation system structured on such results would be of general applicability to the entire population of enlisted naval jobs. The criterion consisted of rankings of the jobs by experienced naval personnel on over-all difficulty and responsibility. The evaluations of the jobs on the various factors were made by Navy job analysts.

## The Samples of Jobs

For the purposes of the investigation the "population" of naval jobs was considered to be those defined in *The Manual of Enlisted Navy Job Classifications* (4).<sup>3</sup>

*Job Dimensions Considered.* In order to obtain representative samples, the following four job dimensions were considered: (1) Job group (14 groups representing different areas, such as quartermaster jobs, electronics jobs, etc.); (2)

<sup>3</sup>The following groups of jobs were excluded from consideration: exclusive emergency service jobs (jobs of relatively restricted scope that typically exist as such only under conditions of full mobilization); jobs applicable to more than one rating; and specialists (job specialties, such as divers, for which some individuals are qualified, and which they may be called upon to perform now and then in addition to their regular duties). The total population of jobs after these exclusions was 825.

<sup>4</sup>*The Manual of Enlisted Navy Job Classifications* classifies jobs into 12 major groups. For the purposes of this investigation, however, they were divided into 14 groups.



Job levels (three levels, namely: Basic, Journeyman, and Supervisory); (3) Branch of service (Aviation versus Non-aviation); and (4) Job location (where the job typically occurs namely: Shore, Shipboard, and Shipboard-Shore).<sup>5</sup>

*Selection of the Two Samples.* The experimental and hold-out samples were then individually so selected that each sample included percentages in each category of the four dimensions which approximated the corresponding percentages in the total population. For the later purpose of deriving criterion values, these two samples were combined, making a total of 103 jobs. Sixteen "extra" jobs were then added to these samples for purposes to be described later, making a total of 119 jobs.

#### *The Criterion*

The "validation" of industrial job evaluation systems usually is in terms of the extent to which a given system results in a satisfactory degree of relationship with prevailing wage or salary levels. Since naval pay for various grades and ranks is established by legislative enactment rather than by labor market "supply and demand" factors, a different type of criterion of job values was necessary. For this purpose twenty-nine representatives of the naval service served as judges in ranking the selected jobs on the basis of over-all difficulty and responsibility.

*Instructions to Judges.* Each judge attended a meeting at which the purposes of the study and the rating procedures were discussed. Each judge was given a packet containing definitions (on separate sheets) of the 119 jobs, and a set of instructions. These instructions asked the judge to select the sheets of those jobs which he felt he could rank in relationship to other jobs, and then to rank these jobs on the basis of "over-all job difficulty and responsibility."

#### *Evaluation of Sample Jobs*

*The Experimental Job Evaluation System.* An experimental job evaluation system with 13 factors was set up for later use in evaluating the sample jobs.<sup>6</sup> Definitions of all factors were incorporated in the experimental system.

<sup>5</sup> This determination was made by analysts in the Bureau of Naval Personnel.

<sup>6</sup> This system included the following factors: (1) Work Knowledge Required\*; (2) Inherent Job Hazards; (3) Guidance and/or Supervision Received\*; (4) Responsibility for Supplies and Equipment\*; (5) Non-hazardous Working Conditions; (6) Physical Effort Required\*; (7) Responsibility for the Safety of Others\*; (8) Guidance, Supervisory and Command Responsibility\*; (9) Potential Combat Hazards and Hardships; (10) Physical Skill\*; (11) Mental Demand; (12) Military and Working Conditions\*; (13) Attention. Eight of these factors (those marked with an asterisk) were the same as those tentatively being considered for use in a service-wide system that was developed in connection with the Military Occupational Classification Program of the Personnel Policy Board, Department of Defense.

*The Job Analysts Who Served as Evaluators.* The jobs were evaluated by experienced job analysts in the naval service. The experimental and holdout samples were evaluated, at different times, by 32 and 13 analysts respectively; 11 job analysts were included in both groups and evaluated both samples.

*Method of Evaluation.* Each analyst was given definitions of the jobs in the sample in question, a set of definitions of the thirteen experimental factors, a set of instructions, and a set of record sheets. The instructions provided for the analyst first to select, from the sample in question, the definition sheets of the jobs which he felt he could rank relative to other jobs on the various factors. Instructions provided then for ranking these selected jobs on each of the 13 factors. For this purpose the "rank-comparison" system described by Bittner and Rundquist (1) was used. This system provides, in general, for the division of the items into subgroups, the ranking of the items within each subgroup, and the subsequent "merging" of the subgroups.

## Results

### *I. Criterion Scale Values*

In order to derive scale values for the 103 original sample jobs as such, the rankings by the judges of the 16 "extra" jobs were disregarded; this was done for each judge by assigning ordinal rank orders to the sample jobs in the order in which he ranked them, exclusive of any of the "extra" jobs which he had also ranked.

*Original Criterion Scale Values.* Since each judge ranked only part of the sample jobs, it was necessary to take this into account in deriving criterion scale values. A method described by Guilford (3, pp. 256-257), appropriate to such situations, was used. These scale values, numerically, ranged from 0 (high) to 3.2037 (low).

*Consistency of Rankings by Judges.* In order to get an estimate of the consistency of the rankings of each judge with the rankings of the entire group of judges, a rank order correlation ( $\rho$ ) was computed for each judge between the rank order of the jobs he ranked and the rank order of those same jobs in the complete array, when the scale values of the jobs he ranked were put into ordinal rank sequence. These rank order correlations ranged from .60 to .94, with a median of .86. While all of these  $\rho$ s were statistically sig-

nificant (at better than the one per cent confidence level), it was decided, in the interests of criterion stability, to drop the rankings of the three judges with the lowest critical ratios. Subsequent analyses were made on the basis of the rankings of the remaining 26 judges; their median  $\rho$  was .87.

While such consistency does not provide demonstrable proof of the validity of the judgments of job values, it lends support for the use of such judgments as a criterion, in the absence of any "true" criterion of naval job values.

*Final Job Samples.* In addition to analyzing the reliability of the criterion rankings among the several judges, an analysis was also made of the consistency with which the individual sample jobs were ranked. While this analysis will not be described in detail, suffice it to say that the 15 jobs that were judged with the least consistency were dropped from the samples, and were replaced, where possible, with "extra" jobs with similar dimension characteristics. Such replacements were not possible for all the jobs dropped, however, and the two samples were reduced to 58 and 37 jobs respectively. The jobs in these two groups (giving a total of 95) were the ones used later in the analysis of the various job evaluation factors.

*Final Criterion Scale Values.* Criterion scale values were then recomputed using these 95 jobs as ranked by the 26 judges mentioned above.

## II. Analysis of Factor Evaluations on Experimental Jobs

*Tentative Rank Orders on Individual Factors.* The factor rankings of the job analysts were used first in deriving tentative rank orders of the 58 experimental jobs on each of the 13 factors. The method presented by Guilford (3, pp. 256-257) for use in deriving scale values from several sets of incomplete rankings involves the intermediate computation of "probability" values. These probability values have the same rank orders as do the final scale values, and were therefore used as the basis for determining these tentative rank orders.

*Reliability of Evaluation by Job Analysts.* The following reliability analysis was made

individually for each of the 32 job analysts. The jobs which each analyst ranked were first extracted from the complete array of the tentative rank orders on each factor; these selected jobs were then assigned ordinal rank orders on each factor in the sequence in which they had been extracted. For each factor a rank order correlation ( $\rho$ ) was computed between the ordinal rank orders of the jobs which the analyst ranked as extracted from the complete array of jobs, and the rank orders of those same jobs as he had ranked them on the factor in question.

The  $\rho$ s for each analyst on the 13 factors were then converted to Fisher's  $z$  values. The 13  $z$  values for each analyst were then averaged, and these averages were then reconverted to  $\rho$  correlations. These average  $\rho$ s ranged from .60 to .89 for the various analysts, with a mean and median for all analysts of .81.

The average  $\rho$ s for the individual analysts were then subjected to a statistical analysis to determine the extent to which they differed from those of the other analysts. The seven analysts whose average  $\rho$ s differed most from those of the remaining analysts were considered as candidates for being dropped for the subsequent analyses. Four of these analysts were dropped. The other three, however, had each ranked certain jobs which in turn had been ranked by limited numbers of other analysts; it was therefore considered desirable to retain the evaluations of these three analysts. The average  $\rho$ s of the 28 analysts retained ranged from .71 to .89, with a mean (computed from Fisher's  $z$  values) of .82. While these values should be considered as being approximations rather than as precise indexes of the reliability of the analysts, the general level of the reliability compares rather favorably with that which is typically obtained in industrial job evaluation studies.

The  $\rho$ s of the 28 analysts for each of the 13 individual factors were then averaged, using Fisher's  $z$  values. These average  $\rho$ s ranged from .64 to .88 for the various factors, with an average  $\rho$  for all factors of .82.

*Final Scale Values on Individual Factors.* The rankings of jobs on the 13 factors by the 28 job analysts were used as the basis for de-



Table 1

Correlations of Factor Scale Values with Criterion Scale Values for Final Experimental Sample

Factor No.	<i>r</i>	Factor Name
1	.954	Work Knowledges Required
2	.193	Inherent Job Hazards
3	-.854	Guidance and/or Supervision Received
4	.631	Responsibility for Supplies and Equipment
5	-.033	Non-hazardous Working Conditions
6	-.139	Physical Effort Required
7	.248	Responsibility for the Safety of Others
8	.629	Guidance, Supervisory and Command Responsibility
9	.048	Potential Combat Hazards and Hardships
10	.420	Physical Skill
11	.756	Mental Demand
12	.089	Military and Working Conditions
13	.505	Attention

giving final scale values for the 58 experimental jobs on each of the factors. The previously mentioned method for developing scale values from several sets of incomplete rankings was used.

*Relationship of Factor Scale Values to Criterion Scale Values.* Table 1 presents the correlations of the scale values on each of the 13 factors with the criterion scale values. These individual correlations range from .954 to minus .854.

In order to determine the factors and their weightings which gave the optimum degree of relationship with the criterion scale values, the Wherry-Doolittle test selection method was used. This method is described by Gar-

rett (2, pp. 435-558). The results of this analysis are given in Table 2. This table shows the factors in the sequence in which they were selected, including the shrunken multiple correlation ( $\bar{R}$ ) with the criterion for each of the selected factors along with all previously selected factors. This table also gives the Beta weights and the subsequently derived "b" weights for the individual factors selected. The first five factors selected gave an  $\bar{R}$  of .968. The addition of the sixth factor caused no increase in the  $\bar{R}$ , indicating that the first five factors by themselves gave the optimum degree of relationship with the criterion scale values. The unshrunk multiple correlation ( $R$ ) of these five factors with the criterion was .970.

It will be observed that Factor 1 (Work Knowledges Required) by itself gave a correlation with the criterion of .954, indicating that this single factor accounted for a very high proportion of the variance in criterion scale values. Factor 3 and Factor 2 entered into the prediction of criterion values with negative weightings. The negative relationship of Factor 3 (Guidance and/or Supervision Received) is readily understandable, but it is interesting to note that this factor "came through" while Factor 8 (Guidance, Supervisory and Command Responsibility) did not. The inverse relationship of Factor 2 (Inherent Job Hazards) is consistent with the typical findings of industrial job evaluation studies.

### III. Cross Validation with Hold-out Jobs

#### Derivation of "Predicted" Criterion Scale Values Using Selected Factors. The five fac-

Table 2  
Shrunken Multiple Correlations with Criterion of Factors in Order of their Selection, with Beta and b Weights, for Final Experimental Sample

Factor Name	Factor No.	$\bar{R}$	Beta Weight	b Weight
Work Knowledges Required	1	.954		.7578
Guidance and/or Supervision Received	3	.963	.7664	-.2505
Potential Combat Hazards and Hardships	9	.964	-.2443	.0631
Inherent Job Hazards	2	.966	.0645	-.1242
Responsibility for the Safety of Others	7	.968	-.1359	.1010
Physical Effort Required	6	.968	.0990	



tors identified as giving the optimum degree of relationship with criterion values of the experimental jobs were used in computing "predicted" criterion values of the 37 hold-out jobs. The first step involved in this process was that of computing scale values for the 37 hold-out jobs on each of the five factors, using the method previously described. The "b" weights for these factors were then incorporated in a regression equation, along with the derived constant ( $K = .2418$ ) in order to obtain the predicted criterion values.

*Correlation between Predicted and Actual Criterion Values.* The predicted criterion scale values for the 37 hold-out jobs were then correlated with their previously determined actual criterion scale values. This correlation was .937. This correlation is of such a magnitude that it gives assurance that the selected factors account for a very substantial proportion of the criterion variance.

### Conclusions

The following conclusions seem warranted on the basis of the results of the investigation:

1. The criterion ranking of the sample jobs by a number of representatives of the naval service reflect fairly stable concepts among them with respect to relative values of enlisted naval jobs; in the absence of any "true" criterion of naval job values, such reliable judgments may well be accepted as a criterion for use in job evaluation research.

2. The rankings of the sample jobs by job analysts on the 13 factors in the experimental job evaluation system resulted in a satisfactory degree of reliability.

3. Five of the 13 factors accounted for a very large proportion of the variance in criterion scale values. For the experimental sample, the shrunken multiple correlation of these five factors with the criterion was .968. For the hold-out sample, the "predicted" criterion scale values (predicted by the use of a regression equation) gave a correlation of .937 with the actual criterion scale values. The third, fourth, and fifth factors identified by the Wherry-Doolittle test selection method (Factors no. 9, 2, and 7) added only slight increments to the shrunken multiple correlation; because of this, it is very probable that the first two identified factors (Factors no. 1 and 3) would themselves adequately predict the criterion scale values, although the predictive value of these two by themselves was not determined in the study.

4. The results of the investigation were of such a nature as to suggest that a job evaluation system structured on the basis of these results could be expected to be of general applicability to the entire population of enlisted naval jobs.

Received August 10, 1953.

### References

1. Bittner, R. H. and Rundquist, E. A. The rank-comparison rating method. *J. appl. Psychol.*, 1950, 34, 171-177.
2. Garrett, H. E. *Statistics in psychology and education* (Third Edition). New York: Longmans, Green and Co., 1947.
3. Guilford, J. P. *Psychometric methods*. New York: McGraw-Hill Book Co., Inc., 1936.
4. *Manual of Enlisted Navy Job Classifications* (NAVPERS 15105, Revised). Bureau of Naval Personnel, Navy Department, July, 1949.

## Comparability of Personal Attitude Scale Administration With Mail Administration With and Without Incentive<sup>1</sup>

Paul W. Maloney

*The Addison Lewis Company, Minneapolis, Minn.*

Industrial psychologists have two obligations. First, the techniques we use must yield valid results. Second, we must be sure that the use of these techniques is economically feasible.

The attitude scale exemplifies an effective device whose utilization has been restricted by cost considerations. Where group administration is possible attitude scales are being used extensively. But many "groups" in the social sense are not grouped geographically. In this latter case the attitude scale is generally administered by personal interview. And that's where the cost per subject zooms upward.

There's no doubt but that mail administration would be cheaper than interviewing. But we have been a little hesitant about administering attitude scales by mail. We have questioned the validity of such an uncontrolled technique.

This study was planned to test validity. We wanted to determine the comparability of three methods of attitude measurements,—one by a personal interview, the other two by mail.

### Method

A total of 127 subjects were personally interviewed. A 19-item Likert-type attitude scale formed part of the questionnaire. The sample was fixed address. Interviewers were permitted to go next door when the second call back was unsuccessful. No records were kept of respondent refusals or the number of next door calls.

A slightly abridged form of the questionnaire used by the interviewers was mailed to 148 subjects. (The attitude scale was not abridged.) Half received a 25¢ piece as in-

centive; this group was sent one blanket follow-up letter. The other 74 subjects were just asked for cooperation; here there were two follow-ups to non-respondents.

### Results

The questionnaire without the quarter received a 58% return. The quarter brought back 86% of the questionnaires.

The same attitude scale had been personally administered to other groups. On the basis of 175 schedules (including the 127 discussed above), the eight most discriminating items were selected. Table 1 shows the average scale values which the three groups made on these eight items.

The two mail techniques produced almost identical average values. Attitudes as denoted by the personal interviews, however, were considerably lower.

As a further test of comparability, two sets of correlations were computed. The first set dealt with the incidence of the median or "Undecided" response. The per cent frequency of this response was computed on each of the 19 items, for all three groups of subjects. These percentages were correlated. The results comprise Table 2.

All relationships were strong. Especially comparable are the two mail surveys. Apparently incidence of "Undecided" does not vary item-by-item among the methods of administration.

Table 1  
Average Scale Values Attained on the Eight Most Discriminating Items

Administration	Average Scale Value
Personal Interview (N = 127)	2.32
Mail, with Quarter (N = 64)	2.58
Mail, without Quarter (N = 43)	2.56

<sup>1</sup> This study was part of a total communications analysis of a public utility. The subjects were all customers of the public utility. The scale concerned customer attitudes toward service, public ownership and company personnel.

Table 2

Correlations of the Incidence of the "Undecided" Response in Three Administrations

Comparison	Pearsonian Correlation
Personal Interview with Mail-Quarter	.84
Personal Interview with Mail-Non-Quarter	.83
Mail-Quarter with Mail-Non-Quarter	.95

The second set of correlations was also designed to test comparability. For each item, the three (out of five) least favorable responses were grouped as non-favorable. The percentages of these non-favorable responses were calculated for each item, for all three groups. The correlations of these percentages are shown in Table 3.

Once again the correlations were high. The two mail techniques were again most similar. Note that the "Undecided" response was one of the three non-favorable. This means that the correlations were to some extent a corollary of the relationships shown in Table 2.

Table 3

Correlation of the Incidence of Non-Favorable Responses in the Three Administrations

Comparison	Pearsonian Correlation
Personal Interview with Mail-Quarter	.84
Personal Interview with Mail-Non-Quarter	.81
Mail-Quarter with Mail-Non-Quarter	.92

## Conclusions

The healthy return indicates a possible economy in mail administration of attitude scales. The per cent of mailed questionnaires returned was influenced by a financial incentive; results were not. Mailed administrations denoted higher attitudes than the personal interview. Though this difference in scale values was pronounced, item-by-item ups and downs were the same with all three types of administration.

Of course, all of these conclusions must be interpreted carefully. Individual attitude scale administrations are specific unto themselves. But if the findings from other studies are similar, we may be able to consider the mailed attitude scale a good tool. The higher scale values are puzzling. But the item-by-item similarity is encouraging. We may find that the technique can be used, providing the denoted attitudes are depressed to some extent. What that depression should be we cannot, of course, say at the present time.

## Summary

Residential customers of a public utility were administered an attitude scale. Three methods of administration were used: personal interview; mail with financial incentive; and mail without financial incentive. The responses obtained by each method were compared. The three were found to be reasonably comparable.

*Received September 3, 1953.*



## An Empirical Analysis of the Effectiveness of Psychological Warfare<sup>1</sup>

Thomas G. Andrews

*University of Maryland*

Denzel D. Smith

*Office of Naval Research*

and

Lessing A. Kahn

*Operations Research Office, The Johns Hopkins University*

Of the many reports on Psychological Warfare (PW) relatively few have been directed toward analyzing hypotheses about the fundamental nature of PW and the ways in which it acts upon the individual recipient. There are, of course, many reasons for this situation. Unfortunately, this is one of the many cases in which we have had to employ a system without full information of the system and how it works. The authors were engaged by the Operations Research Office to design and carry out a research evaluation of several aspects of PW in Korea, especially to determine certain of the antecedent and attendant psychological factors that influence the effectiveness of tactical Psychological Warfare.

As a general working hypothesis, it was assumed that the fundamental effects of PW can be characterized in psychological form, and that they are predictable in terms of the attitudes, motives, and experiences of the recipients. Also, it was hypothesized that PW can affect an individual only in certain optimal conditions. There are, no doubt, wide individual differences in the state of preparedness for the effects of propaganda. Theoretically the individual soldier or civilian who is content with his role, is well taken care of physically, is in no state of fear, and is in

complete accord with the war aims and ideology of the forces of his nation, will not be sensitized to the content of PW. On the other hand, the person who is at the opposite poles of these characteristics may be so ready to surrender or to show defection behavior—or whatever our PW is designed to produce—that he does not need the propaganda. PW was thus thought of as having mainly a nudging or precipitating effect on behavior somewhat secondary to the preparatory effects of the more physical and material aspects of warfare.

### Criteria and Factors Studied

Because the research was to be carried out in Korea and by interrogation of Chinese and North Korean Prisoners of War, the criteria used were restricted by these available conditions. As criteria the research employed the degree of willingness to surrender peacefully as contrasted with having required forceful capture and the degree of disaffection shown.

Attempts were made to identify several factors which would serve as estimates of an individual's position along the general continuum of receptiveness to PW, and to include estimates of behavior that were hypothesized as important conditioners of the criterion behavior. The following nine factors were chosen for the investigation:

A. Degree to which the individual, before the war, was in accord with the ideology and war aims of the Peoples Government.

B. Degree to which, and frequency with which, the individual had experienced intensive fear during battle.

<sup>1</sup> This report was extracted from a more detailed and complete report of the total research project. The material has been approved for presentation at the 1953 meetings of E.P.A. and for publication, the approval being granted by the Operations Research Office and the Department of the Army. The views herein expressed are those of the authors and do not necessarily reflect the opinions of the Army or the Operations Research Office. Official clearance of the material in this report has precluded presentation of several features of the investigation.

C. Degree to which the individual felt he had been poorly treated and physically cared for by his own forces during the war.

D. The amount and intensity of direct battle experience the individual had.

E. The total amount of U. N. Forces propaganda of any kind received by the individual during the war.

F. The total amount of U. N. Forces propaganda per medium received by the individual during the war: (1) leaflets; (2) loudspeaker; and (3) radio.

G. Relative proximity of the propaganda received to action in front line battle.

H. Degree of defection or change in the individual's accord with the aims and operations of his military forces. (criterion)

I. Degree to which the individual was willing to, or sought to, surrender peacefully as opposed to forceful capture at the time he was taken prisoner. (criterion)

Scales were designed to measure the relative position of a person on each of the nine factors. A combination of techniques was used to estimate such positions. Several questions with sets of alternative responses were written for each factor, and the alternative responses were designed in such a way as to reflect a level of intensity or amount of the experience or attitude being assessed. Examples of such items are given below:

A-16. How would you characterize yourself in terms of actions to uphold the principles of the Peoples Government?

- (1) Tried to be critical and show others the faults of communism.
- (2) Was neutral and took no action one way or another.
- (3) Was active in furthering the principles of the Peoples Government.
- (4) Believed so firmly was willing to fight for these principles.

C-34. As the war progressed, to what extent did your living conditions such as food, clothing, comfort, and medical care change?

- (1) Was always able to get along fairly well.
- (2) Things were bad but never unbearable.
- (3) At times things were nearly unbearable.
- (4) Conditions became completely unbearable.

The questions on each factor were grouped together. For the end of each of these sections a 7-point rating scale was designed with descriptive anchors for each point regarding relative position on the scale for the particular factor involved. In addition to these more formalized approaches an open-end interview system was devised for each of the nine factors.

### Procedure

The question forms were translated into Korean and into Chinese and printed in those languages. Through the cooperation of the Army officials in Japan and Korea the authors went to Korea, where a group of native Korean college graduates was selected and trained to serve as interviewers, interpreters, and translators. These men were trained in the procedures of the standardized interview. Military arrangements were made to allow the authors and the native members of the research team into the Prisoner-of-War camps in Pusan, and to furnish groups of POWs selected according to several criteria.<sup>2</sup>

In the interview sessions rapport was established with relative ease, and certain conditions were arranged to assess the veracity of the reports. The interviewer in each case, after explaining the general procedure, read each question and the alternate responses, and checked the response selected by the prisoner. Any seemingly important discussion that was raised about an item was also recorded by the interviewer. At the end of each section of the interview, the rating scale was described in detail and the prisoner indicated his judged position on it. After each rating scale was used, the interviewer discussed the prisoner's experiences in a general way to probe for further comments and descriptions relating to that particular factor. Full notes were recorded on these open-end parts of the interview, and the interviewer made his own rating of the prisoner based on his comments and discussion. This process

<sup>2</sup> The exact nature of the criteria of selection cannot be specified here, nor can the authors describe the Prisoners other than by indicating they were made up of several hundred Chinese and North Koreans captured or surrendering during military operations.

was continued through the schedule of nine factors. The papers were then translated back into English and brought back to the United States for analysis.

A single numerical index was desired for each prisoner on each of the nine factors. A group of research assistants worked on the forms to obtain a single rating on a new nine-point scale for each factor for each of the cases. These new ratings were based on judgments considering all the information recorded. Each form was analyzed independently by at least two assistants. Whenever differences in judgment of final rating value exceeded one scale point, the raters held discussions to resolve the discrepancy. Compromises of one scale position discrepancies were accepted as being sufficiently refined for the data available, and the sets of ratings were averaged.

### Results

The resulting data were processed by IBM equipment and the ratings on the nine factors were intercorrelated. The resulting correlation matrix is presented in Table 1. A correlation above .10 here is significant at the 1% level. One of the general findings of some importance here is the fact that such high correlations were obtained. With data that were suspected to contain relatively large errors of measurement such as these,

the finding of any correlation above .30 was satisfying.

The coefficients in rows H and I indicate the factors that relate to defection behavior and willingness to surrender respectively. Those factors that correlate with one of these two criteria also correlate in the same direction and general magnitude with the other, and the two criterion scales correlate highly with one another. This expected consistency of results for the two criterion scales serves to indicate a core of reliability and credibility of the results for these two scales.

Three of the factors for which scales had been constructed and which had been estimated to have some influence on defection and/or willingness to surrender did not show any significant relation to the two criteria. These particular scales were constructed for estimates of fear (B), amount of PW received by radio (F-3), and the relative proximity of the PW received to operations in front line battle (G). The items on intensity of fear probably did not work for Orientals; they did not correlate well with any other measures. The prisoners reported that they did not have radios available, and so scale F-3 could not be expected to give results.

The morale factors contained in scales A and C correlate higher with the criteria, H and I, than do the Psychological Warfare factors E and F-1. This result was expected,

Table 1  
Obtained Correlations Among Specified Attitudes and Experiences of North Korean and Chinese Prisoners of War \*

	A	B	C	D	E	F-1	F-2	F-3	G	H	I
War aims	A	—									
Fear	B	.02	—								
Bad treatment	C	-.59	.15	—							
Battle exp.	D	.16	.07	-.08	—						
PW rec'd	E	-.18	.02	.24	.40	—					
Leaflets	F-1	-.15	.06	.21	.03	.89	—				
Loudsp'ker	F-2	-.14	-.09	.31	.14	.76	—				
Radio	F-3	-.08	-.10	.03	.19	.52	.09	—			
PW proximity	G	.04	.13	.04	.39	.19	.10	—			
Defection	H	-.58	.03	.52	.33	.28	.31	-.06	—		
Surrender	I	-.59	-.05	.46	.31	.30	.19	.10	.06	—	
				-.25	.20	.22	.08	.06	-.06	.71	—

\* Correlations higher than .100 are significant at the 1% level on a two-tail test.



and it is probably the case in any large-scale military operation. The concern here is the extent to which accuracy of prediction of defection and surrender from amount of PW received is so contaminated with the other factors, such as morale, that one must say that PW bears no significant or demonstrable influence. Attempts were made to analyze this problem by means of partial and multiple correlations.

The net correlation between PW (factor E) and defection (H), partialling out accord with war aims (A), changes the correlation of .31 to .26 which is still significant. When estimates of bad treatment (factor C) are also partialled out, this second order partial correlation of PW and defection (E and H) reduces only to .22, which is still statistically significant. It would seem that Psychological Warfare does offer some effective influence on the Oriental troops, independently of its conjoint action with lowered morale. This net relationship does not, however, appear to hold for predicting relative willingness to surrender peacefully. When the morale factors are partialled out, the predictable effects of PW on surrender behavior reduces to a second-order partial correlation of only .09.

Within the correlation table obtained there are several values that stand out as provocative to consider. Only certain of the relationships are summarized here. Because estimates of chronologically antecedent behavior are being dealt with here, it is more than usually compelling to attribute causality to the results. Analysis of the interrelationships shown in Table 1 appears to indicate that defection and surrender are behavior patterns that are less expected in the more seasoned troops of high morale, but are predictable among green troops of lower morale. This, of course, is practically an established principle based on rationalization and experience in warfare. However, the fact as brought out here serves to demonstrate some reliability of the data obtained and to corroborate the view that morale is a primary target for Psychological Warfare. At least with Orientals, merely reiterating the desirability of surrender and giving suggestions about defection to enemy troops is not enough. It is

also important to note in Table 1 that the morale factors A and C correlate significantly with the total amount of PW received (E). This result may mean either that the PW was destructive of morale or that lowered morale sensitized the troops to the PW.

It was desired to determine whether there are any influences of PW that are independent of the morale determiners of defection and surrender behavior. Through second-order partial correlations one set of results was obtained, as previously described. Further analysis in terms of multiple regression was used to throw light on this problem.

The multiple correlation of factors A, C, D and E with the criterion H, defection, is .66, and the regression equation in Beta-form for this criterion is presented below. The Beta-coefficients serve to indicate relative contribution of the factors mentioned:

$$\begin{aligned} H' &= -.37A' + .23C' - .14D' + .24E' \\ I' &= -.42A' + .15C' - .25D' + .19E' \end{aligned}$$

With the criterion I, willingness to surrender, the multiple correlation with A, C, D and E is .65, and this regression equation in Beta-form is shown above. Comparison of these Beta-coefficients again indicates some of the differential influence of these particular "determining" factors.

When the measure of defection is added to the regression equation predicting surrender and also adding the factor of number of leaflets received, the multiple correlation was .76. This correlation is for the prediction of surrender behavior from a knowledge of all the other important factors. With variables of the type used and obtained under the relatively poor field conditions of measurement that necessarily existed in this study, a multiple correlation of .76 is extremely high. Of course, it contains the contribution of one criterion in predicting the other criterion, in the amount of .71. However, the defection attitudes were presumably antecedent to the surrender or capture of the troops becoming prisoner. In so far as each of the variables other than the scale for surrender is a measure of some behavior occurring prior to the final actual surrender or forceful capture of the prisoners, this correlation of .76 would appear

to indicate that peaceful surrender may be to a great extent dependent on and predictable from the particular forms of attitudes and experiences measured by the scales devised for this study. It is understood that this study requires cross-validation and also replication with other groups.

### Summary

Standardized interviews on North Korean and Chinese prisoners of war were carried out to test the relative importance of several attitudes and experiences in determining the defection attitudes on the part of the captive troops and their willingness to surrender peacefully at the time they were taken prisoner. Among the experiences assessed was the amount of tactical psychological warfare the troops had received from the United Nations before becoming prisoners of war.

"Scores" on each factor and experience were derived as well as on the two criteria of defection and willingness to surrender.

The primary results are presented in a correlation matrix, which is analyzed for certain relations and with respect to the general hypothesis that psychological warfare is effective in changing behavior, but its effects are mainly of a precipitating nature that is differential for persons more sensitized to it by their morale and experiences.

The primary correlations, certain partial correlations, multiple correlations, and standard multiple regression coefficients were analyzed and appeared to corroborate the major hypothesis. Additional relationships of possible military and social importance are deducible from the data obtained.

*Received September 10, 1953.*

## Predicting Achievement in Medical School: A Comparison of Preclinical and Clinical Criteria<sup>1</sup>

Robert Glaser and Owen Jacobs

*American Institute for Research, Pittsburgh, Pa. and University of Pittsburgh*

In a previous article in *J. appl. Psychol.* (3), data were reported on the predictive efficiency of a trial selection test battery administered to an entering class of medical students. The criterion against which the tests were validated was the general grade average at the end of the first year of medical school. Criterion data for later medical school performance have become available and provide the opportunity for a follow-up of this earlier article.

Validation studies of medical aptitude test batteries often employ first-year medical school grades as the criterion variable. The use of these grades as an intermediate criterion of medical success is defensible since the successful completion of the first year is necessary for continuance in medical school and the drop-out rate may be higher during this year than other medical school years. However, the usual question concerning the relationship between an intermediate criterion and an ultimate one still remains. A medical school curriculum can usually be divided into the first two preclinical years and the last two more clinical years. The performance of students in the latter two years, presumably, is generally more similar to their performance as physicians than their performance in the earlier years. It has been pointed out by Stalnaker (5) of the American Association of Medical Colleges that "... the grades given in professional schools may have a special meaning. In the two preclinical years where basic science courses are usually taken, medical schools have one teacher for each 4 to 5 students. In the two clinical years, one teacher is used for each 1 to 2 students. Some schools have more full time teachers (or their equivalent) in the clinical years than they have students. Many—most—of

these clinical teachers are part time and many are voluntary, i.e., unpaid. Grades given under these conditions may have special meaning." The purpose of the present study is twofold: First, to investigate the relationship between preclinical and clinical grades; and secondly, to compare the validities of an aptitude test battery when the criterion variable consists of preclinical grades on the one hand and clinical grades on the other.

This paper reports data obtained from a class of 129 medical students at the Indiana University School of Medicine. At the beginning of the first year of medical school, 150 students were enrolled in this class; at the end of the third year 129 students remained. The two criteria employed were the general grade averages at the ends of the first and third years of medical school. These general averages are weighted averages of the grades obtained by a student in specific courses. Weights are assigned according to the amount of time a course meets. The specific courses in the first and third years are listed in Table 1.

### The Relationship between Criteria

The correlation between the two grade averages is .54. The mean of the first-year averages is 87.4; the standard deviation is 4.1. For the third-year averages the mean is 88.7 and the standard deviation is 2.0.

The correlations of the specific course grades in the first year with the specific course grades in the third year are presented in Table 1. In Table 1 the third-year courses are arranged in order of the size of their mean grades. This makes apparent a noticeable trend in these correlations. As the mean grade decreases, more and more of the correlations between the first- and third-year courses become statistically significant. Of

<sup>1</sup> Based upon a paper presented at the meetings of the Midwestern Psychological Association in Chicago, May, 1953.



Table 1  
Correlations Between First-Year and Third-Year Grades with Means and Standard Deviations of These Grades

Third-Year Courses	First-Year Courses				M	S.D.
	Physiol.	Neuro-Anat.	Histol.	Gross Anat.		
Psychoneurosis	.15	.03	.10	.14	96.3	4.3
G. U. Surgery	.22	.20	.18	.21	95.2	3.7
Ophthalmology	.19	.11	.16	.12	92.6	4.9
Dermatology	.03	-.02	.04	-.10	91.8	2.3
Industrial Med.	.14	.18	.02	.12	89.7	6.0
Epidemiology	.21	.06	.08	.05	89.5	4.4
Clin. Path. Lab.	.31	.39	.20	.29	89.2	4.2
Pathology	.19	.20	.18	.14	88.7	4.2
Clin. Neurol.	.10	.06	.19	-.04	88.2	1.9
Obstetrics	.24	.21	.21	.22	88.0	3.6
Anesthesia	.13	.14	.09	.09	87.2	5.3
Clin. Psych.	.28	.24	.21	.12	87.2	6.8
Clin. Diagnosis	.32	.30	.39	.25	86.7	2.2
Cardiology	.48	.42	.44	.38	86.5	5.3
Pediatric Lect.	.28	.31	.08	-.02	86.2	2.8
Surgical Path.	.26	.35	.40	.30	85.8	5.3
Medicine Recitation	.36	.31	.37	.30	85.6	3.4
Anatomy	.26	.17	.36	.36	83.0	4.3
M	88.5	88.1	86.6	86.3		
S.D.	5.6	4.8	3.8	4.4		

the six third-year courses with the highest means, none correlates significantly with any of the first-year courses. Of the six third-year courses with the lowest mean grades, four of the six correlate significantly with all four of the first-year grades; one of the courses correlates significantly with three of the first-year courses; and one course correlates significantly with two of the first-year courses. No systematic difference in standard deviations, shape of distributions or content of the courses accounts for these results. The means of the first-year course grades are of the same size as the means of the low third-year courses. If this is not an artifact of this sample it may be that when high grades are given for the third-year courses, grading takes place on a different basis than when the mean of the grades is lower.

#### Comparison of Validities for the Two Criteria

A trial aptitude battery was administered to the medical students at the beginning of the first year of medical school. The tests in the battery were the following:

1. The Differential Aptitude Tests—Space Relations (2). This test consists of items which require two-dimensional figures to be translated into their corresponding three-dimensional objects. The rationale behind this test was that an important requirement of the medical student appears to be the ability to translate his two-dimensional text-book illustrations into three-dimensional life objects.
2. The United States Armed Forces Institute Tests of General Educational Development, College Level, Test Three: Interpretation of Reading Materials in the Natural Sciences (6).
3. The Miller Analogies Test (4).
4. The Army General Classification Test, the AGCT (1).

The validities for each test for the first- and third-year general grade averages are given in Table 2. The changes in the coefficients from the first to the third year are not significant.

Table 3 presents the first and third year validities for the medical Professional Aptitude Test (7) which was administered to the students in this class. The changes in the coefficients from the first to the third year are not statistically significant.

For the trial test battery the multiple correlation of the battery with the first-year cri-

Table 2

Intercorrelations, Validity Coefficients, Means and Standard Deviations for the Tests in the Trial Battery

Test	Intercorrelations				Validities		M	S.D.
	1	2	3	4	1st year	3rd year		
1. Reading Interpretation	—	.52	.33	.18	.44	.39	67.3	8.2
2. Miller Analogies		—	.53	.40	.28	.22	60.2	12.1
3. AGCT			—	.31	.04	.13	117.2	9.2
4. Space Relations				—	.16	.13	62.1	15.4

terion is .47, with the third-year criterion .39. The multiple correlation of the best predictors in the Professional Aptitude Test with the first-year criterion is .43, with the third-year criterion .39. (It should be pointed out that some prior selection had taken place on the Professional Aptitude Test upon entrance into medical school. The primary interest in this paper, however, is the change in validity and not the absolute value of the coefficients.) Indication of the overlap between groups when selection is based on each of the two criteria can be obtained by using the regression equation to predict the preclinical and clinical grades for each student and then correlating the two sets of predicted grades. When this is done for the trial battery the correlation coefficient is .40. With this degree of relationship it is possible that, for this kind of test battery, selection based upon preclinical or clinical criteria can make substantial difference in the groups selected.

The results comparing preclinical and clinical criterion grades show no significant changes in test validity. The tests predict

preclinical and clinical achievement equally as well. For the trial test battery, comparison of the predicted scores based upon the two different criteria indicates that the groups selected on the basis of each criterion might be quite different. It can be assumed that achievement in the clinical years is a better indication of performance as a physician than achievement in the preclinical years. If this is the case, then a selection test battery should consist of predictor variables which concentrate upon predicting achievement in the clinical years. Along these lines future test development might well be devoted to the following: (a) the isolation of behaviors which are unique to clinical achievement as compared with preclinical achievement; (b) the development of reliable measures of these criterion behaviors; and (c) the development of testing techniques to predict these behaviors.

Received September 19, 1953.

### References

1. Army General Classification Test, First Civilian Edition. Science Research Associates, 1947.
2. Bennett, G. K., Seashore, H. G., and Wesman, A. G. *Differential Aptitude Tests: Space Relations*. The Psychological Corporation, 1947.
3. Glaser, R. Predicting achievement in Medical School. *J. appl. Psychol.*, 1951, 35, 272-274.
4. Miller, W. S. *Manual for the Miller Analogies Test*. The Psychological Corporation, 1947.
5. Stalnaker, J. M. *Validation of Professional Aptitude Batteries: Tests for Medicine*. Princeton: Educational Testing Service, 1950.
6. United States Armed Forces Institute Tests of General Educational Development, College Level, Test Three: Interpretation of Reading Materials in the Natural Sciences. The American Council on Education, 1943.
7. Vaughn, K. W. *The interpretation and use of the Professional Aptitude Test. A manual for committees on admission in colleges of medicine*. The Graduate Record Office, January, 1947.

Table 3

Correlations between PAT Scores and General Averages and the Means and Standard Deviations of the PAT Scores

Test Score	r		M	S.D.
	1st year	3rd year		
Verbal Ability				
Scientific	.28	.27	562.0	82.2
Social	.15	.18	516.5	75.3
Humanistic	.22	.20	514.3	86.8
Composite	.32	.27	534.5	74.1
Quantitative Ability	.28	.29	541.3	80.5
Index of General Ability	.33	.31	539.2	71.1
Modern Society	.30	.27	533.5	70.5
Premedical Science	.41	.38	561.8	67.5

## Subscore Patterns on ACE Psychological Examination Related to Educational and Occupational Differences

Francis J. Di Vesta

Syracuse University<sup>1</sup>

The present study presents further data on a line of investigation opened up by Munroe (1). In the original study the hypothesis was tested that something of the dynamics of personality might be revealed in the *difference* between the "Q" and "L" scores (Q-L) derived from the American Council on Education Psychological Examination (ACE). The present study is reported here because the findings, derived through pattern analysis as suggested by Munroe, provide implications for personality study and test interpretation.

### Background

The rationale behind the hypothesis presented by Munroe was based upon studies of scatter analysis, particularly those of Rapaport (2) on the Wechsler-Bellevue scale. Her procedure was to administer ACE and the Rorschach tests to 80 students. The Rorschach responses were then related to the *difference* between the individuals' "Q" and "L" percentile standings. On the basis of whether the "Q" or "L" percentile score was the higher, subjects were classified into two groups, one called the *higher Q* group and the other called the *higher L* group. The Rorschach entries for each group were then analyzed for differences.

There were found to be: (a) significantly more V entries (lack of accurate form) for the *higher L* group; (b) significantly more F entries (responses in which form was the determinant) for the *higher Q* group; and significantly more M (movement) entries for the *higher L* group. Description of the two groups based on differences in Rorschach indices might be as follows: The *higher Q* group gave responses which show objectivity through elaboration by careful observations

of objective details, formal intellectual approach, repressive efforts at control of affect and inhibition of normal creative imagination and of normal structuring of perception.

The *higher L* group gave responses which show subjectivity and imagination through cues serving as springboards to new ideas, lack of objectivity, creative organization and a subjective approach, sometimes to excess. *It should be noted that Munroe did not verify the findings for either group by cross validation.*

Roe (3) indicates that the syndrome represented by the *higher Q* group is similar to that found in paleontologists as reflected in their Rorschach protocols. The presentation of this fact was further supported by citing evidence from an unpublished study by Munroe wherein it was found that in a sample of college students the *higher Q* group tend to choose more scientific and art subjects while in college than does the *higher L* group.

### Statement of the Problem

The summary of findings presented above would seem to indicate that the Q-L constellation (or pattern) may be related to differences in the utilization of intelligence arising from differences in personality (4). However, since Munroe's findings were based on groups at the extremes of a continuum (the top and bottom quartiles of the distribution of Q-L scores) and since they were a unique sample, it appeared desirable to test the original findings with evidence from other populations.

Specifically, it is intended in this study, to examine the findings that *high Q* and *high L* scores reflect different personality syndromes through empirical demonstrations of the relationship of these scores to occupational and curricular selections made by the individual.

<sup>1</sup> This study was conducted while the author was employed by the Human Resources Research Institute, Maxwell Air Force Base, Alabama.



## Procedure

ACE was administered to subjects in an advanced military school. The average age of the sample was 32 years of age. They averaged two years of college education, and were professionally well established.

In calculating the difference between the "Q" and "L" scores a minor variation of the procedure used by Munroe was made. This variation was made in determining the relative standing of the individuals through the use of the standard score rather than in terms of the percentile standing. Standard scores for the "Q" and "L" scores were computed on the basis of the distribution of the groups to which the ACE tests were administered.

All calculations in the study were based on the scores of the entire population except in individual cases where complete data were not available. Where groups were dichotomized on the Q-L variable all cases which had a difference greater than zero, regardless of how small the difference, were placed in either the *high Q* or *high L* categories.

## Results

The first step in the analysis was to determine the relationship between the ACE "Q," "L," total and Q-L scores for the population studied. These relationships are shown in Table 1.

The major hypothesis was that the Q-L pattern is related to the kinds of occupational and curricular choices made by individuals. A sub-hypothesis was formulated that pilots would have different Q-L patterns than non-pilots. The rationale was that whether an individual becomes a pilot is initially a matter of choice although the non-pilot population may not be as homogeneous with respect

Table 1

Intercorrelations of ACE Scores

ACE Score	ACE Score		
	"L"	Total	Q-L
"Q"	.57	.81	.50
"L"		.94	-.40
Total			-.08

Table 2

Q-L Scores Obtained by Flying and Ground Personnel

		Acro-Rating					
		Flying Pilots		Flying Non-Pilots		Ground	
Q-L Score	N	N	Per Cent	N	Per Cent	N	Per Cent
<i>Higher Q</i>	220	154	57	16	43	50	36
<i>Higher L</i>	226	116	43	21	57	89	64
Total	446	270	100	37	100	139	100

$$\chi^2 = 16.86; N = 2; p = .001.$$

to choice. Thus, the pilot population may be considered to represent a group of individuals who expect to utilize their intelligence and skills in a specified manner. Accordingly, because the demands of the pilot's job are highly technical and require a high degree of objectivity it was expected that more pilots would be in the *higher Q* group than non-pilots. Because the non-pilots are in administrative positions it was hypothesized that they would tend to be in the *higher L* group.

The number of pilots, flying officers other than pilots (navigators, bombardiers and observers) and ground officers in the *higher Q* and *higher L* groups are shown in Table 2. The chi-square for this distribution is significant ( $p < .01$ ). The greatest contribution to chi-square was found in the difference between the numbers of the non-flying individual in the *higher Q* and *higher L* groups, although each of these classifications contributed to the total chi-square.

Since data on another population of about 400 more personnel were available, it was decided to duplicate the first analysis as a check. Trends in this second population were as distinct as the ones shown in Table 2. The chi-square was lower but was significant at the .02 level of probability. The consistency in the two populations is attributable to the fact that the major contribution to chi-square is made by the differences between the number of non-flying personnel in the *higher Q* and *higher L* categories. F ratios for the ACE "Q," "L" and total scores were not significant for these populations.

A second sub-hypothesis was made that there would be a difference between the number of reserve officers and regular officers in the *higher Q* and *higher L* categories. The rationale was that the regular officers represented a homogeneous group of individuals who selected the Air Force as a career, whereas, the reserve officers represented a more heterogeneous group of individuals with respect to making a career in the Air Force. In the first sample there was a tendency for more regular officers to be located in the *higher Q* category than in the *higher L* category and more reserve officers to be located in the *higher L* than in the *higher Q* category. The chi-square for differences in these distributions was not significant ( $p < .20 > .10$ ). The same tendency existed in the second population but again the difference in the distributions was not significant ( $p < .30 > .20$ ) by the chi-square test. These findings indicate that the Q-L constellation does not differentiate between these two groups.

A third sub-hypothesis was that there would be differences between groups of individuals in different areas of greatest job experience. This hypothesis, as in the cases of the other sub-hypotheses, was selected because the area of job experience appeared to represent the kind of occupational activity in which the individual was most interested. Consequently, the inference was made that the demands of the job areas involving planning (e.g., comptroller) would be filled by individuals with *higher L* scores and job areas involving more mechanical or technical routine requirements would be filled by individuals with *higher Q* scores.

The means and standard deviations of the Q-L scores for individuals in each job area are shown in Table 3. The F ratio for the data in this table is 2.07 ( $p < .05 > .01$ ). The data clearly indicate that the Q-L scores differentiate between individuals in the maintenance-inspection type of function and the individuals in the intelligence-comptroller type of function. Those in the former functions tend to have *higher Q* scores and those in the latter functions tend to have *higher L* scores. Means for the second population were calculated, and the F test was again applied.

The F ratio for the variances between groups in the second population was found to be significant. In addition, the F ratio for variances between the two populations was also found to be significant. Differences between the populations were found to be attributable to the Q-L scores for individuals in the maintenance and operations areas. The mean for the maintenance subjects in the new population was  $49.1 \pm 1.6$  and the mean for the operations subjects was  $47.9 \pm .8$ .

The original "Q" and total scores of the ACE failed to discriminate between subjects in the various job areas. The F ratio for each of these sets of scores was significant at the  $> .05$  level of probability. The F of 2.1, however, was significant ( $p < .05$ ) for the original "L" scores of subjects in these job areas. In decreasing order from high "L" scores to low "L" scores, as shown in Table 4, were these job areas: intelligence, research, communications, comptroller, supply, personnel, administration, operations, inspection and maintenance. There is some tendency for the ACE "L" score to discriminate between the job areas for this population in the same order that the ACE Q-L score does. There is, however, enough difference to indicate that something different is being measured by the Q-L pattern from that being measured by the "L" score.

Table 3  
Mean Q-L Scores of Individuals in Ten Occupational Areas

Occupational Area	N	Mean	Standard Error of Mean
Maintenance	37	53.9	1.2
Inspection	9	51.2	2.0
Communications	8	51.0	2.8
Research	8	50.9	3.2
Operations	138	50.3	0.7
Personnel	35	49.1	1.4
Administration	67	48.7	1.1
Supply	12	48.5	2.6
Intelligence	24	45.0	1.6
Comptroller	9	44.1	2.6
All other	44	49.4	1.4
Unknown	5	45.6	1.3

$$F = 2.1; p = < .05 > .01.$$

Table 4

ACE Total and Subscore Means of Personnel in Ten Occupational Areas

Occupational Area	N	"L" Score		"Q" Score		Total Score	
		M	S.E.	M	S.E.	M	S.E.
Maintenance	37	67.4	2.3	40.8	1.5	108.2	3.5
Inspection	9	68.8	5.2	39.2	2.1	108.0	6.9
Communications	8	77.7	5.7	43.5	3.0	121.3	7.9
Research	8	80.8	4.4	45.0	2.3	125.8	5.0
Operations	138	75.2	1.5	41.5	0.8	116.6	2.1
Personnel	35	75.6	2.7	40.7	1.5	116.3	3.8
Administration	67	75.5	2.1	40.1	1.3	115.6	3.1
Supply	12	76.7	3.4	40.6	2.2	117.3	4.5
Intelligence	24	84.8	3.0	41.6	1.9	126.4	4.4
Comptroller	9	77.2	5.7	36.7	3.2	113.9	8.2
All other	44	78.2	2.4	42.9	1.4	121.1	3.1
Unknown	5	86.4	4.6	42.0	3.9	128.4	8.3
F =		2.1		0.6		1.4	
p =		<.05>.01		>.05		>.05	

A final hypothesis was that the Q-L pattern would discriminate between individuals who chose different fields of specialization in college. This variable was selected because it too represented a situation in which choice was a factor. Fields of specialization in college were grouped into five classifications for purposes of analysis. The *science* group was represented by such college majors as psychology, geology, mathematics and zoology; the *arts* group by majors in liberal arts, history, music and English literature; the *technical* group by majors in the applied areas (excluding engineering) such as education, law, social casework and agriculture. The *engineering* and *business administration* groups were composed of majors in these respective fields of specialization.

The percentage of subjects in the *higher Q* group and in the *higher L* group for each of the five types of college majors is shown in Table 5. The differences in proportions of *higher Q* and *higher L* subjects in these classes of college majors is significant ( $p < .02$  and  $> .01$ ) by the chi-square test. These findings do not support those of Munroe and Roe who found in their studies that students selecting science subjects tended to have *higher Q* than *higher L* scores.

The F ratios between fields of specializa-

tion for the "Q," "L," and total ACE scores were significant ( $p < .01$ ). These are shown in Table 6.

Students who had specialized in science, arts, and engineering achieved higher scores on the ACE "L" score than did those who majored in technical and business administration courses. Students who majored in technical courses achieved lower scores on the ACE "Q" scores than did any of the other groups. Majors in engineering, arts and science achieved higher total scores than did those with business administration and technical fields of specialization.

Table 5

Proportions of Students in Each Field of Specialization with *Higher Q* and *Higher L* Scores

College Major	N	Per Cent in <i>Higher Q</i> Group	Per Cent in <i>Higher L</i> Group
Arts	83	43	57
Sciences	100	48	52
Technical	74	54	46
Engineer	94	57	43
Business Administration	66	70	30
Total	418		

$\chi^2 = 12.41$ ;  $N = 4$ ;  $p = <.02>.01$ .



Table 6  
ACE Total and Subscore Means for Personnel with Different College Majors

College Major	N	"L" Score		"Q" Score		Total Score	
		M	S.E.	M	S.E.	M	S.E.
Science	100	78.8	1.6	45.2	0.9	124.1	2.2
Arts	83	80.1	1.8	46.2	1.0	125.8	2.4
Technical	74	73.0	1.9	42.4	1.0	115.5	2.7
Engineer	94	78.3	1.7	47.8	1.0	126.0	2.5
Business Administration	66	72.0	1.8	45.4	1.0	117.4	2.6
F Ratio		3.96		3.75		3.71	
p		<.01		<.01		<.01	

### Summary

The present study was conducted to examine the findings of Munroe that the *high Q* and *high L* patterns (derived from ACE sub-scores) reflect different personality syndromes. To demonstrate whether the original findings would apply in a different research situation the Q-L scores were studied in relationship to occupational and educational differences.

A sample of Air Force officers was used in the present study. The criteria used were as follows: the rating (flying or ground) of the officer; the officer's assignment to the regular or reserve corps; the officer's career field; and the officer's college major. The findings were that: pilots tended to have *higher Q* scores, non-pilots had *higher L* scores; individuals in maintenance and comptroller jobs had *higher L* scores; personnel with college majors in arts and sciences had *higher L* scores whereas individuals in the applied areas, engineering and business administration had *higher Q* scores. No difference was found in the Q-L patterns of reservists and regular officers.

Although these data indicate a relationship between the Q-L pattern and occupational and educational choices there is a question as to whether this pattern represents a predisposing factor or whether it emerges as a result of experience in certain areas.

The original ACE scores were not found to be consistently related to these situations. No differences were found between the ACE

"Q," "L" and total scores of flying and ground personnel. The ACE "Q" and total scores did not discriminate between individuals in different job areas although the "L" score did discriminate in about the same order as the Q-L pattern. Each of the original ACE scores was related to the college majors but not in the same manner as was the Q-L pattern score.

It would appear from this evidence that there is a relationship between the ACE Q-L pattern and the utilization of intelligence by the individual. As further studies are performed using one or another form of pattern representation, it is reasonable to expect that more will be revealed through test score patterns than through independent scores or summation of these scores. Certainly the use of a clinical approach to the understanding of personality dynamics underlying pattern representations must be considered a useful first step in the development of hypotheses for empirical investigations.

Received August 10, 1953.

### References

1. Munroe, Ruth L. Rorschach findings on college students having different constellations of sub-scores on the ACE. *J. consult. Psychol.*, 1946, 10, 301-316.
2. Rapaport, D. *Diagnostic psychological testing. Vol. I Menninger Clinic Monograph Series No. 3.* Chicago: Year Book Publishers, 1945.
3. Roe, Anne A. Rorschach study of a group of scientists and technicians. *J. consult. Psychol.*, 1946, 10, 317-327.
4. Super, D. E. *Appraising vocational fitness.* New York: Harper and Brothers, 1949.

# The Effect of Methods of Presentation and Examining Conditions on Student Achievement in a Correspondence Course

Francis J. Di Vesta

Syracuse University \*

Correspondence courses form an important communication medium for education in both military and civilian instructional areas. These courses may take the form of highly commercialized businesses independent of an educational institution, home study courses conducted by schools, or extension courses supported and administered by military and other governmental agencies. The number of students enrolled in such courses probably numbers 100,000 or more per year. Despite the importance of this particular segment of our educational system a review of the literature reveals a most obvious scarcity of studies applying directly to effective methods for the presentation of these courses. Accordingly, when the question concerning the most effective method of presenting correspondence courses occurred in a military extension course institute a ready answer was not available in the literature. Consequently, it was decided to conduct an experiment in which the correspondence course students would be used as the experimental population. The present report is a summary of the findings from this experiment.

## The Problem

The problem in this study was to determine the most adequate methods of presenting course materials for effective student achievement. Two major hypotheses were the specific foci of the experiment. The first was that three styles of presenting correspondence course materials would result in differential student achievement. The second hypothesis was that quality control, as imposed by examining conditions, would affect

the achievement of the students and their retention of level of achievement.

## Procedure

The study was conducted with applicants enrolling for a physical training course. This course was at the Officer Candidate School level of difficulty and was devoted to an understanding of the development of physical education programs for combat fitness.

The manual or text in the course was prepared in three different "styles." *Style A* was a manual written in a popular and personal manner with several illustrations of the cartoon variety. This manual was commonly referred to as the *Popular Science* style for descriptive purposes. *Style B* was written in the formal expository manner commonly used in textbooks. Detailed illustrations were used in the text manual. *Style C* was actually a study guide divided into several "lessons" or units. Each unit had its major objective(s), references and questions. An Air Force field manual was provided with the study guide for reference purposes. This style was known as the "Chicago" style because it was generally fashioned after the syllabi used in the University of Chicago home study courses.

The research design is described briefly below.

1. All enrollees were administered, through the mails and before receiving course materials, a pre-test of fifty items.
2. Each enrollee was then assigned to one of the following experimental groups:

Kind of Examination	Style of Material		
	Style A	Style B	Style C
Open book examination	Style A	Style B	Style C
Closed book examination	Style A	Style B	Style C

Assignment to these groups was made on the basis of pre-test scores so that equal numbers of students in each quartile would appear in each of the experimental groups.

3. Enrollees in the groups taking the open book examination received the "final" examination at the time they received the course materials to complete in any manner they desired.

Enrollees in the groups taking the closed book examination named a proctor. When the indi-

\* This study was conducted in the Officer Education Division of the Human Resources Research Institute, Maxwell Air Force Base, Alabama. Dr. Lora McDonald, Extension Course Institute of the Air Force, provided useful guidance and assistance in the design and administration of the experiment.

vidual felt he was ready to take the examination, he reported to the proctor who, in turn, administered the examination. The completed examination was returned by the proctor, with certification, to the administration center.

Course materials were returned by both groups at the time the examinations were returned to the administration center.

4. Thirty days after taking the final examination, students took the same examination a second time. This test was known as the "retention" examination. Enrollees in both open and closed book examination groups took the retention examination without course materials. The "open book" group took the retention examination without a proctor and the "closed book" group under the same proctorship as in the original administration.

5. When the retention examination was received by the administrative office the course materials were returned to the student for his files.

6. Enrollees were notified at the start of the study that they were to participate in a research, that they would be notified of each step in the research *only* before the time it was to occur and that they would be informed when the research was completed.

The course was administered through regular administrative mailing procedures. A total of 900 enlisted airmen enrolled in the course over a three month period. Eight months later a total of 353 individuals completed all steps required in the research. The results reported here are based on this group.

### Results

The means for each of the groups on the pre-test<sup>1</sup> are shown in Table 1. Although enrollees were assigned in equal numbers to each of the experimental groups, more stu-

Table 1

Pre-Test Scores for Each of the Experimental Groups

Style*	Closed Book Exam			Open Book Exam		
	N	M	S.D.	N	M	S.D.
A	49	32.4	3.9	68	32.3	4.7
B	46	32.9	3.5	73	31.7	4.1
C	51	31.9	3.7	66	31.3	2.1

F (open vs. closed book) = 2.2;  $p = >.05$ . F (between styles) = 1.6;  $p = >.05$ .

\* See text for description.

<sup>1</sup> The pre-test was an examination used for previous classes of enrollees. A new examination was used for the "final" examination in the experiment.

Table 2

Final Examination Scores for the Experimental Groups

Style*	Closed Book Exam			Open Book Exam		
	N	M	S.D.	N	M	S.D.
A	49	42.5	5.8	68	48.8	6.8
B	46	44.2	7.3	73	50.1	6.8
C	51	45.2	8.7	67	48.3	8.5

F (between styles) = 2.8;  $p = >.05$ . t (closed vs. open book) = 6.5;  $p = <.01$ .

\* See text for description.

dents in the open book examination groups completed the course than did those in the closed book examination groups.

The variance between the "open book" and "closed book" subjects' performance on the pre-test was not significant ( $p > .05$ ) as revealed by the F of 2.22. The F of 1.59 was not significant ( $p > .05$ ) for the variance between pre-test scores for groups of subjects assigned each type of material. On the basis of these data, and in the absence of more definite information about the subjects' characteristics it was assumed that the groups were from the same population.

The second step in the analysis was to determine differences in performance of each of the experimental groups on the final examination.<sup>2</sup> The mean achievement scores and standard deviations are shown in Table 2 for each of the experimental groups.

The difference between the achievement of subjects who were assigned course material written in different styles was not significant. Subjects taking the open-book examination, however, achieved significantly ( $p < .01$ ) higher scores than those taking the closed-book examination ( $t = 6.54$ ).

A similar analysis was made for the retention test results. The retention test was in every respect the same test that was given for the final examination. It was given thirty days after receipt of the final examination

<sup>2</sup> The final examination was pre-tested on three groups: individuals who had taken the course, individuals in physical training and individuals who had neither taken the course nor had physical training experience. Items were selected on the basis of discrimination function and on reliability. The Kuder-Richardson reliability coefficient of the total instrument was .88.



Table 3  
Retention Examination Scores for the  
Experimental Groups

Style*	Closed Book Exam			Open Book Exam		
	N	M	S.D.	N	M	S.D.
A	47	42.3	5.9	67	45.0	7.1
B	45	41.0	12.7	70	45.6	8.7
C	51	44.5	8.6	67	45.7	8.3

F (between styles) = 0.9;  $p = >.05$ . t (open vs. closed book) = 3.0;  $p = .01$ .

\* See text for descriptions.

and was administered in the absence of formal course materials. Again the style in which the course was written appeared to have no significant effect on the retention of the students' achievement level. The quality control did, however, have an effect. The achievement scores of individuals who took the closed-book examination were significantly ( $t = 2.96$ ) lower than those who took the open-book examination. The means and standard deviations for these groups are shown in Table 3.

The average individual loss between the final and retention examination and the extent to which the experimental variables were a factor in retention of achievement level is shown in Table 4. Differences in losses were not found to be significant for the types or styles of materials. The losses in achieve-

Table 4

Losses During the Time Between the Final Examination and the Retention Examination for Each of the Experimental Groups

Style*	Closed Book Exam			Open Book Exam		
	N	M	S.D.	N	M	S.D.
A	47	-0.3	4.9	67	-3.9	4.6
B	42	-0.5	4.6	69	-3.8	5.6
C	51	-0.7	3.6	67	-2.6	5.7

F (between styles) = 0.8;  $p = >.05$ . t (open vs. closed book) = 5.8;  $p = <.01$ .

\* See text for description.

ment level of those subjects who took the closed-book examination, on the other hand, were significantly less ( $t = 5.67$ ) than the losses in achievement of subjects who took the open-book examination.

### Summary

The present study is a report of an experiment conducted with a correspondence course population. Two hypotheses were investigated during the course of the experiment. One hypothesis was that three different styles of presenting course materials would have differential effects on student achievement and retention of achievement level. The second hypothesis was that the degree of quality control, as imposed by the open and closed book examination, would have no effect on the achievement level and retention of the achievement level by students.

The styles of presenting course materials (popular, expository, and study guide) were not found to be different in their relative effectiveness as measured by an achievement examination. Nor did these methods affect the retention of the achievement level. On the other hand, the subjects who used the examination with reference to course materials (open book examination) had higher final and retention examination scores than did those students who took the examination under monitorship without the use of the text materials (closed book examination). The subjects who took the closed book examination maintained their original achievement level while those who took the open book examination made significant losses over a thirty-day period.

The administrative procedures required in conducting the experiment with correspondence course populations were found to be too ponderous for practical purposes. The recommendation is made that hypotheses be tested with more readily available populations of subjects and the results applied to correspondence course usage.

Received August 21, 1953.

## The Use of Levels of Confidence in Item Analysis

Valentine Appel<sup>1</sup>

*Richardson, Bellows, Henry & Co., Inc.*

and

David Kipnis<sup>2</sup>

*New York University*

One of the problems of item analysis is the standard to be employed in selecting items for inclusion in a final scoring key.<sup>3</sup> Typically, some level of confidence is arbitrarily chosen and those items which discriminate at this level are selected. Guilford (6) as well as others have indicated that the 5% and 1% levels of confidence should be used as guides when selecting items.

Little consideration has been given in the literature to the influence of the size of the item analysis sample upon the level of confidence at which any given item is likely to discriminate. Most writers have assumed the availability of large samples, and so this problem has probably not been considered particularly important. Large samples, however, are often impossible to obtain, particularly in applied research. The test constructor is often placed in a situation where if any item analysis is to be employed at all, it must be performed on a small sample, frequently less than 100 cases.

Item validities, when computed against an external criterion, are typically low. Given a small item analysis sample, the resulting expected item validities often cannot be reasonably expected to exceed levels of confidence as rigorous as the conventional 1% and 5% levels. The establishment of such rigorous standards would therefore be expected to result in the rejection of a large number of truly discriminating items.

This was recently demonstrated in a study

by Feldman (4), although within a rather narrow range of levels of confidence. He used the 1%, 2%, and 5% levels as standards for item inclusion with high and low criterion groups of 42 cases each. On cross validation, he found that the key containing only those items which discriminated at the 1% level was generally less valid than the keys containing all items which discriminated at the less rigorous 2% and 5% levels.

It would be expected that, given a fairly large item analysis sample, more of the items will show validity exceeding a rigorous level of confidence. The establishment of a less rigorous standard would therefore be more likely to result in a greater proportion of nonvalid than valid variance being added to the scoring key. The problem becomes one of striking the most favorable balance between the number of truly valid items rejected and the number of truly nonvalid items selected.

The purpose of this study was to test the general point that an important consideration in establishing a standard for item inclusion is the size of the item analysis sample available. More specifically, the hypothesis was tested that for maximal test validity, the smaller the sample size available, the less rigorous should be the level of confidence selected as a standard for item inclusion. Conversely, given a large sample size, maximal validity can be achieved by establishing a more rigorous standard.

### Method

*Instruments and Population Employed.* The study was performed using the RBH Supervisory Judgment Test (SJT) to predict an intelligence test criterion provided by the short form of the Armed Forces Qualification Test (AFQT).

The SJT is a test which has been found useful for the prediction of supervisory success. It con-

<sup>1</sup> Presently with Nowland & Schladermundt, Greenwich, Conn.

<sup>2</sup> Formerly with Richardson, Bellows, Henry & Co., Inc.

<sup>3</sup> Although item difficulty and item intercorrelation also represent significant problems in this area, this paper shall concern itself only with the problem of item validity.

sists of 33 items of four and five alternatives in which the examinee is presented with a series of supervisory problems and asked to choose the best and worst alternative for solving each.

The AFQT is a 25-minute timed intelligence test consisting of 45 items covering the areas of verbal, mathematical, and spatial abilities.

As a result of a supervisory selection study which had been recently completed (5), a number of cases were available in which examinees had been administered both the SJT and the AFQT. The experimental population consisted of 540 first, second, and third line civilian supervisors at two United States Army Arsenals.

**Item Analysis.** For purposes of item analysis the experimental population was randomly divided into three samples of 80, 150, and 300 cases. The remaining 10 cases of the 540 were not included in the item analysis. They were later included, however, in the validation series. For each of these samples the following procedure was followed: High and low criterion groups were designated by selecting the upper and lower 27% of the AFQT distribution. The percentage of cases in the high and low criterion groups who responded to each SJT alternative was then determined and the significance of the difference between the percentages in the two groups was computed.

**Construction of Scoring Keys.** From the item analysis data derived from each of the three samples, four plus and minus unit weighted scoring keys were constructed for predicting the AFQT criterion. These keys were composed of all item alternatives discriminating at and beyond the 1%, 5%, 20%, and 50% levels of confidence, one key being constructed for each of these four confidence levels. A total of 12 scoring keys were constructed in all. The number of scored alternatives comprising each of these keys is summarized in Table 1.

**Validation of Constructed Keys.** To validate the scoring keys which were developed on the basis of the item analysis, it was essential to employ samples independent of the samples from which the keys had been developed. In order to fulfill this requirement and also to make maximal

Table 2  
Validity Coefficients for the Twelve Keys

Item Analysis Sample Size	Group*	Level of Confidence			
		50%	20%	5%	1%
80	A	.664	.576	.596	.501
	B	.676	.611	.563	.516
	C	.617	.636	.523	.392
	Mean	.653	.608	.561	.471
150	D	.699	.730	.702	.634
	E	.651	.604	.528	.496
	F	.677	.689	.670	.715
	Mean	.676	.677	.639	.623
300	G	.711	.735	.714	.700
	H	.647	.654	.651	.605
	I	.585	.612	.614	.605
	Mean	.650	.670	.661	.639

\* Each group is composed of 60 cases.

use of the available data, a procedure was followed similar to one recently proposed by Katzell (7).

The cases employed in each of the item analysis samples were systematically reassigned so that the scoring keys constructed from one item analysis sample were employed to score cases selected from the other samples. Thus, for example, the cases which were employed in the item analysis of the 300 case sample were systematically redistributed to form groups which could be used to score the keys which were developed from the 80 and 150 case samples. The cases from the 80 and 150 case samples were similarly reassigned.

Following this procedure, nine independent validation groups (designated A through I), each containing 60 cases, were formed. Groups A, B, and C were assigned to be scored with the four scoring keys developed from the 80 case item analysis sample; Groups D, E, and F were assigned to be scored with the four scoring keys developed from the 150 case item analysis sample; and Groups G, H, and I were scored with the four keys developed from the 300 case item analysis sample. The product-moment correlations of each of the four scores with the AFQT criterion were then computed for each of these 60 case validation groups.

### Results

The validity coefficients computed for each of the keys on each of the validation groups are summarized in Table 2. To test the hy-

Table 1

Number of Scored Alternatives in Each Key \*

Level of Confidence	Item Analysis Sample Size		
	80	150	300
1%	8	27	55
5%	34	52	96
20%	82	110	143
50%	167	187	221

\* The total possible number of scored alternatives, including "best" and "worst" responses, was 302.



Table 3  
Analysis of Variance of the Validity Coefficients

Source of Variation	Sum of Squares	df	M <sup>2</sup>	F
Between sample sizes	.1266	2	.0633	1.92
Between groups of same size	.1974	6	.0329	
Total between groups	.3240	8		
Between levels of confidence	.0903	3	.0301	3.42
Level of confidence $\times$ sample size	.0528	6	.0088	2.75*
Pooled groups $\times$ level of confidence	.0580	18	.0032	
Total within groups	.2011	27		
Total	.5251	35		

\* Significant at the 5% level.

pothesis that the differences among the validities of the various keys could be attributed to errors of sampling, an analysis of variance of the validity coefficients was carried out. Since it is known that the sampling distribution of  $r$ 's does not meet the assumption of normality required for analysis of variance, each  $r$  was transformed to its  $z'$  equivalent, the distribution of which is known to be normal (3). The analysis of variance was carried out according to the Type I design outlined by Lindquist (8) and also discussed by Edwards (2, Chap. 15). The results of this analysis have been summarized in Table 3.

The error terms employed in this analysis bear some discussion. Since the variance between sample sizes for item analysis was based upon coefficients computed from independent samples, the error term which was employed was the variance between groups of the same size. The variance attributable to the interaction between level of confidence and sample size, however, was not based upon estimates derived from independent samples, the four coefficients in any row of Table 2 being based upon the same cases. The error term which was used in this instance was therefore the pooled interaction terms for groups of the same sample size by level of confidence. When tested against this error term, the variance attributable to the interaction between level of confidence and sample size was significant beyond the 5% level. The variance attributable to this interaction was therefore employed as the error term in

testing the significance of the level of confidence main effect.

Only the variance attributable to the interaction between level of confidence and sample size was significant beyond the 5% level. Neither of the main effects were statistically significant. This may be interpreted to mean that, within the limits of the present study, there is no one optimal level of confidence to be employed as a standard for item inclusion. Rather, the optimal level of confidence is a function of the sample size employed for item analysis.

Examination of Table 2 would indicate that the smaller the sample available for item analysis, the less rigorous should be the level of confidence employed. In short, the hypothesis tested was essentially substantiated.

### Discussion

The results of this study, insofar as they may be generalized, indicate that there is no one optimal level of confidence which should be employed when item analyzing test data. Particularly pertinent is the result that such arbitrarily designated confidence levels as the 1% and 5% often cannot be expected to result in maximal cross validities. In many cases, particularly when the size of the sample available for item analysis is small, a much less stringent standard may be expected to result in higher validities than the more conventional 1% or 5% levels.

Especially striking is the fact that, for all sample sizes employed in this study, the 50%

scoring keys consistently resulted in higher validities than the keys composed of items which discriminated at the 1% level. It should be noted, however, that the validities of the 50% key based upon the 300 case sample had started to shrink although the 5% and 1% keys showed continuous increments in validity as the item analysis sample sizes were increased. This would suggest that if larger samples had been employed in the item analyses the greatest validities would have been produced by the 5% and 1% keys.

It would appear that any arbitrarily chosen level of confidence is likely to be a poor standard for item inclusion. Levels of confidence, if they are to be employed at all, ought to consider the sample size available for the item analysis. The smaller the sample size, the less rigorous should be the level of confidence required.

It would seem that standards for item inclusion might profitably be established without any reference to levels of confidence. That is, instead of specifying in advance that only items discriminating at the, say, 1% or 5% levels be included in the test, an alternate procedure is suggested. Such a procedure would entail the computation of item validity indices which are independent of the sample size upon which the item analysis is based, e.g., biserial  $r$ , phi coefficient, etc. These items would then be arranged in decreasing order of validity and a cutting point selected above which items would be selected for inclusion in the scoring key and below which they would be discarded. Since few principles are available as to where the optimal cutting point should be, the decision as to what constitutes minimally acceptable item validity will probably have to be an arbitrary one based upon the judgment of the test constructor.

Only after the items have been selected should any reference be made to the level of confidence at which they discriminate. The level of confidence corresponding to the minimally acceptable standard of item validity can then be determined, and the number of items exceeding this standard can be compared with chance expectancies (1). If the selected number of item alternatives exceeds chance expectancy, it is likely that a scoring key composed of these items will continue to discriminate if applied to new samples.

Received August 31, 1953.

### References

1. Brozek, J. and Tiede, K. Reliable and questionable significance in a series of statistical tests. *Psychol. Bull.*, 1952, 49, 339-341.
2. Edwards, A. L. *Experimental design in psychological research*. New York: Rinehart & Co., 1950.
3. Ely, J. H. Studies in item analysis 2: Effects of various methods upon test reliability. *J. appl. Psychol.*, 1951, 35, 194-203.
4. Feldman, M. J. The effects of the size of criterion groups and the level of significance in selecting test items on the validity of tests. *Educ. Psychol. Measmt.*, 1953, 13, 273-279.
5. PRB Research Note 12, Edgerton, H. A. and Thomson, K. F., et al. *Development of Techniques for the Selection of Wage Board Supervisors at Army Arsenals*, 30 June 1953. Copies of this report may be obtained from the American Documentation Institute, Auxiliary Publications Project, Photoduplication Service, Library of Congress, Washington 25, D. C. Order Document Number 4092: Microfilm copy, \$3.50; photostat copy, \$10.00.
6. Guilford, J. P. The phi coefficient and chi-square as indices of item validity. *Psychometrika*, 1941, 6, 11-19.
7. Katzell, R. A. Cross validation of item analyses. *Educ. Psychol. Measmt.*, 1951, 11, 16-22.
8. Lindquist, E. F. *Design and analysis of experiments in psychology and education*. Boston: Houghton Mifflin Co., 1953.

## Some Computational Short-Cuts in the Development or Analysis of Tests

Angus G. MacLean and Arthur T. Tait

*California Test Bureau, Los Angeles, California*

In developing a new test or in evaluating an existing one, the procedure is to administer it to a sample of the population for which it is intended, and obtain the following statistics: Mean; Variance; Reliability; Item difficulties; and Item-test correlations.

In addition to these statistics, since the selection of items on the basis of item-test correlations does not insure homogeneity of test content (2), some index of item-value is needed. In a recent article (1) the present authors suggested: (a) computing the inter-item covariance total for each item; and (b) selecting only those items whose covariance totals exceed their variances. The former indicates the contribution of each item to reliability (homogeneity) while the latter points up the contribution to unreliability (heterogeneity).

In the same article a method was described whereby all the information listed above, plus the *S*-indices (item-selection indices), can be obtained in one operation, and with less computation than is required for computing each of these statistics separately and by the usual formulas. Accuracy is also improved and duplication of computations eliminated. For the sake of exhibiting the rationale and general method, the full computational procedure was given. However, there are many short-cuts and, as the specific purpose varies, some steps may become unnecessary. At this point it is proposed to describe the most economical methods of: (a) effecting the preliminary arrangement of the data; (b) obtaining the mean, variance (standard deviation) and reliability; (c) obtaining the item difficulties and/or the mean item difficulty; (d) obtaining the item-test correlations; and (e) obtaining the selection-indices. These elements are arranged sequentially, i.e., each earlier step is necessary to each subsequent one. Items are assumed to be scored 1 or 0, and 'reliability' refers to Kuder-Richardson formula 20.

(a) A count is made (by hand or machine), of the number of cases "passing" each item<sup>1</sup> denoted by  $f_{ii}$ , and of the number of cases passing both of every possible pair of items, denoted by  $f_{ij}$ . If there are  $n$  items there will be  $n$   $f_{ii}$ 's and  $\frac{n(n-1)}{2}$   $f_{ij}$ 's. These frequencies

are then displayed in the *F*-matrix, consisting of an  $n \times n$  table. In row 1 column 1 is placed  $f_{11}$ . In row 2 column 2 is placed  $f_{22}$ . The diagonal from top left to bottom right is called the principal diagonal and its elements, if divided by  $N$ , the total number of cases, become the item "difficulties," usually written  $p_{ii}$ , or proportion passing on each item. In row 1 (corresponding to item 1), all other cell entries indicate the number of cases passing on both item 1 and the item corresponding to the column number. Thus in row 1 cell 5 the entry will be the number of cases who gave the correct response to both items 1 and 5. This is denoted by  $f_{15}$ . Once every subject's response to every item has been punched, the IBM electronic statistical machine can generate the *F*-matrix very quickly. If there are enough counters, it saves computing time to obtain the sum of the entries in each row at this time.

(b) Once the *F*-matrix (which should, by the way, be symmetrical, i.e.,  $f_{ij} = f_{ji}$ ) is obtained, there is little computing to do. First, if this is not already accomplished by machine, obtain the row sums, and then obtain  $T$ , the sum of these. Thus  $T$  is the sum of all the entries in the matrix. A quicker way to compute  $T$ , if only the mean, variance (S.D.) and reliability are desired, is to sum the frequencies on one side of the principal diagonal only, double this sum and add  $\sum f_{ii}$ , which must be obtained separately anyway. It helps to avoid mistakes if a line is ruled along the principal

<sup>1</sup> When neither a tabulator nor an IBM Electronic Statistical Machine is available, it is best not to construct an *F*-matrix, but to adopt a procedure which will be described below.



diagonal, from the top left corner of the matrix to the bottom right.

It is also necessary to obtain  $\sum f_{ii}$  and  $\sum f_{ii}^2$ ; these are computed simultaneously on modern desk calculators.

Operations of the type  $kx - yz$  or  $kx - y^2$  are single operations on modern calculators, so that a computational entity has come into being known as  $L$ , defined as  $N^2\sigma^2$ ; that is, if  $L$  is divided by  $N^2$  the variance is obtained.  $L$  is defined by:

$$L_{xx} = N\sum X^2 - (\sum X)^2. \quad (1)$$

In the Kuder-Richardson formula 20,

$$r_{ii} = \frac{n}{n-1} \left( 1 - \frac{\sum \sigma_{ii}^2}{\sigma_i^2} \right), \quad (2)$$

where  $\sigma_{ii}^2$ , the variance of item  $i$ , is defined by

$$\sigma_{ii}^2 = p_{ii} - p_i^2. \quad (3)$$

But it is unnecessary to divide both numerator and denominator in (2) by  $N^2$ , so

$$r_{ii} = \frac{n}{n-1} \left( 1 - \frac{\sum L_{ii}}{L_{ii}} \right), \quad (4)$$

where  $\sum L_{ii} = N\sum f_{ii} - \sum f_{ii}^2$ ,

and  $L_{ii} = NT - (\sum f_{ii})^2$ .

Then  $\sigma_i^2 = \frac{L_{ii}}{N^2}$ , (5)

and  $M_i = \frac{\sum f_{ii}}{N}$ , (6)

where  $\sigma_i^2$  denotes the variance of the test and  $M_i$  the mean.

In (2) the factor  $\frac{n}{n-1}$  is present because  $\sigma_h^2$ , the sum of the inter-item covariances, or  $(\sigma_i^2 - \sum \sigma_{ii}^2)$ , is not precisely the true variance,  $\sigma_\infty^2$ , but

$$\sigma_\infty^2 = \frac{n}{n-1} \cdot \sigma_h^2. \quad (7)$$

Therefore, since reliability is defined as the ratio of true to total variance, or that proportion of the variance which is not attributable to error,

$$r_{ii} = \frac{\sigma_\infty^2}{\sigma_i^2} = \frac{n}{n-1} \cdot \frac{\sigma_h^2}{\sigma_i^2}. \quad (8)$$

The square of the standard error of measurement is defined by

$$\sigma_e^2 = \sigma_i^2 - \sigma_\infty^2 \quad (9)$$

$$= \sigma_i^2 - \frac{n}{n-1} \cdot \sigma_h^2, \quad (10)$$

but the usual formula, which is equally efficient *provided* enough significant figures are used for  $r_{ii}$ , is

$$\sigma_e^2 = \sigma_i^2(1 - r_{ii}), \quad (11)$$

which, incidentally, demonstrates that

$$\sigma_\infty^2 = r_{ii}\sigma_i^2. \quad (12)$$

It may be of interest to note that (5) may be written

$$\begin{aligned} \sigma_i^2 &= \frac{NT}{N^2} - \frac{(\sum f_{ii})^2}{N^2} \\ &= \frac{T}{N} - M_i^2. \end{aligned} \quad (13)$$

(c) If the separate item "difficulties" are required, e.g., to arrange the items in order of difficulty, obtain  $\frac{1}{N}$  to sufficient significant figures and lock it in as a constant multiplier, then run down the principal diagonal converting each  $f_{ii}$  in turn into  $p_{ii}$ , i.e.,

$$p_{ii} = \frac{f_{ii}}{N}. \quad (14)$$

If on the other hand only the mean difficulty is required

$$\bar{p} = \frac{1}{nN} \sum f_{ii} = \frac{M_i}{n}. \quad (15)$$

Should the variance of item difficulties be desired, it is clearly obtained by

$$\sigma_p^2 = \frac{n\sum f_{ii}^2 - (\sum f_{ii})^2}{n^2N^2}. \quad (15a)$$

(d) If item-test correlations (and/or selection indices) are required, the sum of each row in the  $F$ -matrix should have been recorded, e.g., in a column to the right of the matrix. Then first obtain  $\sigma_{ii}$ , the item-test covariance, as follows:

$$\sigma_{ii} = \frac{1}{N} \left( \sum f_{ri} - \frac{f_{ii}}{N} \cdot \sum f_{ii} \right), \quad (16)$$

where  $\sum f_{ri}$  denotes the row-total for item  $i$ . If difficulties have been recorded in a column,

$f_{ii}$  can be substituted for  $\frac{f_{ii}}{N}$ . In (16)  $\sum f_{ii}$  is, of course, a constant and it is simpler from a computational point of view to make two columns out of (16), first locking in  $\sum f_{ii}$  and obtaining the differences and then multiplying them by the locked-in reciprocal of  $N$ . If, however, the items are to be selected by use of the  $S$ -index rather than by selecting the best  $r_{ii}$ 's, do not use (16) at all, but see section (e) below.

The item-test correlations are obtained by:

$$r_{ii} = \frac{\sigma_{ii}^2}{\sqrt{\sigma_{ii}^2 \cdot \sigma_i^2}}, \quad (17)$$

where  $\sigma_{ii}^2$  and  $\sigma_i^2$  are already available. This is the point-biserial correlation between item and total test.

(e) The selection index for item  $i$ ,  $S_i$ , is defined by

$$S_i = \sigma_{ii}^2 - 2\sigma_{ii}^2. \quad (18)$$

[For the full explanation see reference (1).] If  $S_i$  is negative, item  $i$  should be rejected, the more so, the greater its absolute value. However, if  $\sigma_{ii}^2$  has not been computed,  $S_i$  is defined by

$$S_i = \frac{1}{N^2} (L_{ii} - 2L_{ii}). \quad (19)$$

The best computational procedure is to obtain, for each item, the value  $(N\sum f_{ri} - f_{ii} \cdot \sum f_{ii})$  and record it; then obtain each  $(Nf_{ii} - f_{ii}^2)$  and record it. It will be noticed that the latter is  $L_{ii}$  while the former is  $L_{ii}$ . Then subtract  $2L_{ii}$  from  $L_{ii}$  and record the difference, with algebraic sign. Finally, multiply each difference by  $\frac{1}{N^2}$ . This last step would not be

necessary if the purpose were to reject *all* items with a negative  $S$ , whatever the magnitude. However, it has been found in practice that if the substantially negative items are eliminated, so that their rows and columns are removed, those with small original negative  $S$ 's may now have only positive covariances with the remaining items, and their covariance totals may now exceed their variances, giving them a positive  $S$ -index. Therefore, the large negatives should be eliminated first and the indices re-

calculated. Also, as  $n$  diminishes, the increase in the ratio  $\frac{n}{n-1}$  sometimes more than compensates for the increase to be gained in  $r_{ii}$  by rejecting one or two small negatives.

#### Procedure When Machine Facilities are Limited

Some of the statistics used in the above formulas are identical with those arrived at in the ordinary process of obtaining individual scores and their mean and standard deviation.  $T$  is the quantity usually denoted by  $\sum X_i^2$ , that is, the sum of the squares of the individuals' total-test scores, and  $\sum f_{ii}$  is  $\sum X_i$ . The only other statistics needed are  $\sum f_{ii}^2$  and  $\sum f_{ri}$ . With regard to the first, it is customary, in developing a test, to obtain the frequency passing each item in order to compute the item difficulties, so nothing unusual is called for. The only extra step required by this method is obtaining, for each item, the quantity  $\sum f_{ri}$ .

Now, as long as items are scored 1 or 0,  $\sum f_{ri}$  is the same as  $\sum X_{ii}$ , the sum of the total-test scores of those who gave the correct response to item  $i$ . This is a cross-product sum and gives rise to the formula

$$L_{ii} = N\sum X_{ii} - f_{ii} \cdot \sum X_i. \quad (20)$$

$F_{ii}$  is, of course, identical with  $\sum X_{ii}$ . The other formulas above may similarly be rewritten, substituting  $\sum X_i^2$  for  $T$ ,  $\sum X_i$  for  $\sum f_{ii}$ , and  $\sum X_{ii}$  for  $\sum f_{ri}$ .

The most elementary way to obtain  $\sum X_{ii}$  would be to hand-sort the answer-sheets. It is quicker and easier (and has other advantages, as will be seen) either to punch item-scores and total-test scores on IBM cards and use a mechanical sorter, or to use "needle-sort" cards, which are punched around the edge by hand and sorted by inserting a sorting needle. This last method was recently tried out so as to determine the amount of time it would consume. Total scores were written on the cards, and a sort was made for each item, then the scores of those who had gotten the item right were added on a desk calculator. There were 124 items and 100 individuals (1 card for each), and the whole operation, from punching to the final summing of the  $\sum X_{ii}$ 's as a check, took between 19 and 20 hours, or about 9 minutes

per item. This is an over-estimate of the general average because, with that number of items, two cards had to be stapled together for each individual, and they had to be very carefully fitted together so that the holes would be in the right positions. With 100 or fewer items this time-consuming step would be eliminated and the sorts would be quicker too; about 7 minutes per item would be a fair allowance with  $N = 100$ . With a key-punch and a mechanical sorter the punching and sorting would be quicker, but the real saving in time occurs when a tabulator is available for the adding—here two or two and a half minutes per item is ample allowance.

The full  $F$ -matrix method has, however, certain advantages, the most prominent of which is this: if any items have to be eliminated, all that is required is to delete the corresponding rows and columns and to obtain new row sums. If one is using  $\sum X_{iu}$ , on the other hand, the answer-sheets have to be rescored and new total-test scores punched, and the sorting-and-summing operation repeated. Thus, the  $F$ -matrix, though more work at first, is likely to be less work on the whole unless the number of items is large. Of course, if there is no question of eliminating items, but only of eval-

uating the merits (or demerits) of the items present, the  $\sum X_{iu}$  method is the more economical, the greater the number of items involved. The reason for this is that the time it requires increases linearly with the number of items, while the time required by the  $F$ -matrix increases as the square—thirty-five minutes for a 15-item test, about 4 hours for a 50-item test, and probably two days for a 100-item test. If the matrix can be broken down into subtests of not more than 15 items each, however, the time required is again a linear function—roughly half an hour for each such subtest. If this cannot be done, it is better to use the  $\sum X_{iu}$  method, eliminate all items with negative  $S$ -indices, and repeat the scoring, sorting and summing to obtain the final item values.

Received April 5, 1954.

Early publication.

#### References

1. MacLean, A. G. and Tait, A. T. A procedure for analyzing a test and maximizing its reliability. *J. exper. Educ.*, 1954, 22, 3.
2. Mosier, C. I. A note on item analysis and the criterion of internal consistency. *Psychometrika*, 1936, 1, 275-282.



## Some Relationships Between the MMPI and a Problem Checklist

Robert F. Lockman

*Student Counseling Bureau, University of Minnesota*

The purpose of this study was to determine what, if any, relationships exist between a problem checklist used in the Student Counseling Bureau at the University of Minnesota and the Minnesota Multiphasic Personality Inventory. Berdie (1) related this same problem checklist to the Minnesota Personality Scale and found in his sample that students with low scores (indicative of the presence of problems) on various sections of the Scale tended to indicate related problems on the checklist. An added purpose of the present study was to compare checklist responses with those in Berdie's study conducted eight years earlier.

The problem checklist<sup>1</sup> contains 33 items and instructs the student to check those which he has not adequately solved and to double check those which he wants to discuss with a counselor.

### Procedure

Checklist responses and MMPI T-scores were obtained for 335 men and 125 women students counseled at the Bureau during the 1948-1949 college year. This sample included all college and pre-college students for whom complete data were available. Students with MMPI's of doubtful validity were eliminated. Cutting scores of 70 on the L scale, 80 on the F scale, and 60 on the ? scale were used as the validity criteria (2).

### Results

*Checklist Analysis.* The most frequently checked items (single and double checks combined) dealt with educational and vocational problems. Over 80 per cent of the men and 70 per cent of the women indicated that they were unable to determine what they were best able to do; over 50 per cent of both sexes did not know what they wanted to do. One reason for these results may be the fact that the

Student Counseling Bureau is primarily an educational and vocational guidance center. Other frequently checked problems were concerned with job opportunities, duties, and training requirements and with study habits.

In the personal-social problem area, over 30 per cent of both sexes felt that they lacked self-confidence. Twenty-five per cent of the women felt that they did not have enough to talk about in social situations.

Investigation of single and double checking of the more frequently expressed problems indicated that the subjects seemed more willing to discuss their educational-vocational problems with a counselor than their personal-social problems. They may have perceived the Student Counseling Bureau mainly as a place to obtain help on these kinds of problems. Educational and vocational problems are probably more socially acceptable and personally admissible than those dealing with personality and social relations. In attempting to explain this phenomenon, Berdie states that: "Reluctance to discuss certain types of problems may be due to the fact that the students think that nothing can be done about (them). . . . They may consider their personal problems too private to discuss with a relative stranger. . . . When students come to the counselor, they come with one primary purpose and all other matters may appear irrelevant at that time."

On the checklist, a significant difference (.05 level of confidence) existed between men and women on only one item: I have been unable to determine what I am best able to do. Approximately 82 per cent of the men checked it, while only 70 per cent of the women did so. Otherwise, the men and women were roughly equal on relative percentages of problems checked. It is not known whether this is due to the structure of the checklist, the actual incidence of such problems in these groups, or other unidentified variables.

<sup>1</sup> A reproduction of the checklist may be found in Berdie (1).

**Comparison of Checklist Responses.** Comparison of the total percentages of checks in Berdie's study with the present investigation yielded no significant differences between the women in these two samples. However, seven significant differences were found on checklist problems between the male samples. Significantly greater percentages were found by Berdie on two items:<sup>2</sup>

I usually feel inferior to my associates (.05)

I do not know how to obtain the money I need (.05)

In the present study significantly greater percentages were found on the following items:

I am unable to determine what I would like to do (.05)

I am frequently embarrassed when with others (.05)

I have so much outside work that I am neglecting my school work (.05)

<sup>2</sup>The numbers in parentheses following the item indicate the level of confidence.

I do not know how to take good lecture notes (.01)

I am not interested in my studies (.01)

The above differences may be a function of sample sizes and composition (e.g., the loading in the present study of returned servicemen), an actual change in student problems over a period of time, or of other factors not readily apparent.

**MMPI Characteristics of Groups Checking Many and Few Problems.** The median total number of problems checked, regardless of their nature, was four for the men and five for the women. The male average was 4.8 with a standard deviation of 22.9; the female average was 4.9 with a standard deviation of 12.4. Thus, the men were nearly twice as variable in the sheer number of problems they checked than were the women. On the basis of these statistics, the male and female groups were separated into two groups: the "High" group (checking five-or-more problems) and the "Low" group (checking four-or-less problems). Critical ratios were com-

Table 1  
Comparisons of Mean T-Scores of High and Low Problem Groups

Comparisons of Mean T-Scores of High and Low												
MMPI Scale	Men						Women					
	Low		High		CR	$r_b$	Low		High		CR	$r_b$
	(N = 190)		(N = 145)				(N = 63)		(N = 62)			
	Mean	S.D.	Mean	S.D.			Mean	S.D.	Mean	S.D.		
?	50.1	0.4	50.0	0.0	—	—	50.0	0.0	50.1	0.4	—	—
L	52.0	4.1	51.0	2.9	2.61*	.17°	52.2	4.0	51.7	4.4	0.71	.08
F	53.3	4.6	55.8	7.0	3.84**	-.27°°	52.3	4.9	54.5	6.6	2.11*	-.23°
K	57.8	8.6	52.4	9.2	5.49**	.37°°	58.4	8.4	52.6	8.7	3.82**	.41°°
Hs	51.0	7.3	51.6	8.3	0.66	-.05	49.0	5.0	48.9	7.7	0.70	.01
D	51.1	9.6	55.5	11.9	3.63**	-.25°°	50.2	7.8	52.2	8.1	1.43	-.16
Hy	55.4	7.3	55.9	8.3	0.56	-.04	53.9	6.8	53.7	9.2	0.14	.02
Pd	57.4	8.5	59.5	10.2	2.04*	-.14°	54.4	10.7	56.9	11.3	1.28	-.14
Mf	56.7	10.7	58.6	11.2	1.57	-.11	47.8	13.1	50.6	8.5	1.41	-.16
Pa	51.6	7.2	53.9	9.2	2.48*	-.18°	53.0	8.5	56.6	8.6	2.38*	-.26°
Pt	56.2	8.3	60.8	11.5	4.05**	-.28°°	52.8	8.0	56.4	9.3	2.32*	-.26°
Sc	55.8	8.2	60.4	12.0	3.97**	-.28°°	54.5	7.1	58.2	10.3	2.31*	-.26°
Ma	57.5	10.3	59.6	10.5	1.83	-.13	55.0	11.1	57.4	11.8	1.18	-.13
IE	47.4	8.7	53.0	10.0	5.46**	-.37°°	50.0	7.5	53.4	10.9	2.06*	-.23°

\* Significant at the .05 level of confidence.

\*\* Significant at the .01 level of confidence.

° Significant difference from zero at the .05 level of confidence.

°° Significant difference from zero at the .01 level of confidence.

puted between the High and Low groups on each MMPI scale. The results are presented in Table 1.

Both the male and female High groups were significantly higher than the Low groups on the F, Pa, Pt, Sc, and IE scales. The F scale indicates "faking bad" or inability to comprehend the inventory items. The Pa scale indicates tendencies toward sensitivity, hostility, and difficulty in taking criticism. The Pt scale indicates tendencies toward anxiety, indecisiveness, and feelings of inadequacy and insecurity. The Sc scale indicates tendencies toward fantasy, shyness, and withdrawal. The IE scale indicates tendencies toward social introversion (2).

The male High group was also significantly higher than the Low group on the D scale (indicating depression, discouragement, or rejection of a situational or prevailing nature) and the Pd scale (indicating nonconformity, irresponsibility, impulsiveness, and asociality).

Both the male and female Low groups were significantly higher than the High groups on the K scale. This scale indicates test-consciousness, defensiveness, and an attitude of problem denial. The male Low group was also significantly higher than the male High group on the L scale, a measure of the degree to which the subject may be attempting to falsify his scores by always choosing the response that puts him in the most socially acceptable light.

Biserial  $r$ 's (see Table 1) for all of the above comparisons were significantly greater than zero, but the amount of overlap of the High and Low groups was too great to enable accurate classification into these groupings solely on the basis of MMPI scores alone. Nor would the number of problems an individual checked be effective in predicting his MMPI scores. The significant differences obtained, then, are chiefly statistical rather than practical in nature. Only tendencies for these groups may be legitimately pointed out on the basis of these differences. It does seem, however, that individuals who check many problems in this sample tend to have somewhat more deviant MMPI profiles than those who check few problems, although those who check few problems may be denying the existence of other difficulties (high K score).

*Checklist Responses of Subjects Grouped According to Their Highest MMPI Scale Score.* Another method of treating the data was to group the men and women separately according to their highest score on the MMPI clinical scales. An individual's highest scale score would be the one indicated by the highest "peak" on his MMPI profile, regardless of score magnitude. For both men and women, approximately 50 per cent of each group checked problems 6 and 10 on the checklist. These were the most frequently checked items for the whole sample, so they are valueless as far as differential prediction is concerned.

Half of the men with highest scores on the D, Pt, and IE scales indicated on the checklist that they lacked self-confidence. In other words, there was a tendency in this sample for an admitted lack of self-confidence to accompany characteristics assessed by the MMPI as depression, anxiety, indecision, compulsiveness, feelings of inadequacy and insecurity, and withdrawal tendencies.

Half of the men whose highest score was on the Pa scale checked problems related to a lack of job information and reading difficulties. High Pa scores are interpreted as indicative of sensitive, hostile, and paranoid tendencies.

Half of the women with highest scores on the Sc and IE scales stated on the checklist that they did not have enough to talk about in company. Sc and IE peaks are indicative of shy, withdrawing, socially introverted behavior. Half of the women with Pa peaks also indicated that they lack job information as did half of the men with the same highest MMPI score.

In general, there seems to be some logical correspondence between several of the checklist problems and personality characteristics as assessed by the MMPI. This relationship is more obvious for the D, Pt, IE, and Sc scales than it is for the Pa scale.

Since the number of individuals in most of the highest scale groups was so small (median  $N$  for women = 5; for men = 23), valid inference from the above results is impossible. The data should be considered only as descriptive of the sample employed and as a stimulus for further research. It is conceiv-



able that with sufficiently large homogeneous MMPI scale groups, differential problem syndromes might be found on the checklist. Pattern analysis of both the checklist and the MMPI (3, 4) and their interrelations might also prove to be a fruitful technique. The value of such research would be in obtaining stable correlates of personality with respect to expressed problems and stated needs as indicated by the problem checklist.

### Summary

Analyses of the problem checklist and its relations to the MMPI showed that:

1. The most frequently checked problems dealt with educational and vocational difficulties.

2. Men students were nearly twice as variable in the number of problems they checked as were the women students, although the average number of problems checked by each sex was roughly the same.

3. Over a period of time, the relative percentages of responses on the checklist items did not appreciably change for the two samples compared.

4. The subjects seemed initially less reluctant to discuss recognized educational-vocational problems than recognized personal-social problems with a counselor.

5. Both men and women students who checked five-or-more problems on the checklist (as opposed to those who checked four-or-less) had statistically, though not practically, significant higher mean scores on the

F, Pa, Pt, Sc, and IE scales and significantly lower scores on the K scale of the MMPI. Men students checking five-or-more problems also had significantly higher Pd and D scores and significantly lower scores on the L scale. Biserial  $r$ 's for all of the above comparisons were significantly greater than zero.

6. Aside from the most frequently checked problems in the whole sample, half of the men students with MMPI peaks on D, Pt, and IE felt that they lacked self-confidence; half of the women students with Sc and IE peaks felt that they did not have enough to talk about in company; half of both men and women with Pa peaks indicated a lack of job information, while these men also checked problems dealing with reading difficulties. Extreme caution is needed in generalizing from these results since the criterion groups were too small in most instances for stability or validity of results derived from them. These data, then, should be considered merely as descriptive.

Received August 21, 1953.

### References

1. Berdie, R. F. An aid to student counselors. *Educ. psychol. Measmt.*, 1942, 3, 281-290.
2. Hathaway, S. R. and McKinley, J. C. *Manual for the MMPI*. New York: Psychological Corp., 1945.
3. Hathaway, S. R. and Meehl, P. E. *An atlas for the clinical use of the MMPI*. Minneapolis: University of Minnesota Press, 1951.
4. Meehl, P. E. Configural scoring. *J. consult. Psychol.*, 1950, 14, 165-171.

## Facilitating Legislative Research

Harry A. Grace

*Michigan State College*<sup>1</sup>

Legislative behavior has been of periodic interest to many psychologists. Two methods of analysis have been used. A small sample of issues is selected and legislators compared according to their votes on these topics (1, 10, 15). Or a few legislators have been studied on many topics (5, 6, 7). Such restricted studies concentrate on a few legislators and a small number of topics. The basic paradigm for these legislative studies does not differ radically from the familiar sociometric analyses of industrial and social psychologists (2, 8).

A major limitation has been the difficulty of tabulating joint voting (9). Associated with this weakness are other shortcomings. Reliability studies of voting are almost nonexistent (4). Data are presented in tabular form and thus relationships among these data remain vague (16, 18, 20, 22).

This paper reports a method for rapid tabulation of such data. In final form the data are in a symmetric matrix to which a variety of statistics may be applied.

### Procedure

The official legislative journals provide the records from which data are obtained. Information in these records describes the men, their districts, the issues upon which they vote, and the roll call votes they cast. Thus, in our analyses we may control for the legis-

lative body, time of meeting, topics, chairman, etc.

The data are transcribed on standard mark-sensing IBM cards. This process is rapid. The card accommodates up to 54 simple items of data. More than one card may be used to transcribe larger legislative bodies. The roll calls list men alphabetically, and so men are assigned to columns on the card in alphabetical order. The content of the topic on which the vote is taken may also be coded on the card, as may other control information. If each vote is coded in chronological order, easy reference may be made to the journal to check discrepancies. All data are marked on the card by electrographic pencil.

One card is used for each type of vote. If only split roll call votes are tabulated, this means a minimum of two cards (affirmative and negative), and a maximum of four cards (affirmative, negative, absent, abstain) for each vote. A 1 is marked in a man's column on the card which represents the type of vote he has cast. If he does not vote one way, he votes another. Therefore, he will have a 1 in *one and only one* card for each vote. The other cards for that vote will be blank in his column.

Punching the cards is accomplished by machine. It is advantageous for comparative-historical analyses to have the data arranged in a definite, permanent order. A suggested order is numerical, according to the number of the district represented by the legislator, with the First District in column one, and so forth. Thus, if men should fail at the polls, retire, or die, the position of the district representative is unchanged. When the reproducing punch is wired for mark-sensing, the data may be rearranged from the alphabetical order of the men to the numerical order of the district.

The cards are prepared for checking by sorting them on the basis of the vote numbers. The accounting machine (Type 402)

<sup>1</sup> Dr. Gloria Lauer Grace assisted in the design of these studies. The studies have been financed by the University Research Board, University of Illinois, 1950-1952, and the All-College Research Committee, Michigan State College, 1952-1953. Leonard P. Staugas, Statistical Service Unit, University of Illinois, designed the wiring of the accounting machine, Types 402-403. Victor E. Buys, Supervisor of Tabulating Operations, Statistical Methods Section, Division of Disease Control, Records, and Statistics, Michigan State Department of Health, designed the wiring of the electronic statistical machine, Type 101. Norma E. Taschner, Tabulating Office, Michigan State College, and Doris L. Duxbury, Statistical Methods Section, Michigan State Department of Health, were most cooperative in permitting the use of their IBM facilities.

is wired for addition, printing a minor program total each time the vote number changes. Each vote is listed with its content code, the identification of the legislature, and the total of all the cards for each vote is printed. See Table 1. If the mark-sensing and punching are correct, a series of 1's appears in the columns representing the legislators. If a zero appears, it means that the legislator has been overlooked. If a 2 or greater appears, the man has been given credit for having cast more than one type of vote on an issue. Correction of these errors may be made by referral to the pencil markings on the cards. If the cards have been incorrectly marked, reference must be made to the journal. The method is remarkably accurate. The importance of having a machine check rather than a hand check cannot be overestimated in accuracy and amount of time saved. Should subdecks for controls on time or content be reproduced from the master deck, it is highly advantageous that these also be machine checked. The investigator is then assured of a perfect working deck at all times. If errors later appear, he knows that they are a function of the machine operations and not the cards.

The final process is the tabulation of the *joint-occurrence matrix*.<sup>2</sup> Two methods are possible. Either the accounting machine (Type 402-403) or the electronic statistical machine (Type 101) may be used. The accounting machine takes about four times as long and is liable to greater error than the electronic statistical machine. The essential

Table 1

Facsimile of the Verification of Voting Data  
(Columns 1-20 represent legislators; a zero [0] or two [2] indicates an error for that man on the particular vote.)

Vote Number	Legislators																			
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
3	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1
5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

<sup>2</sup> The *joint-occurrence matrix* will be referred to as the *jo-matrix*.

Table 2

Facsimile of the Joint-Occurrence Matrix  
(Column 1 and row 1 identify the legislator from district one, etc. The diagonal is constant, showing the number of times each man voted. The other cells indicate the number of times each man has voted with every other man. The symmetry of the matrix indicates that the data are correctly tabulated.)

Legislators	Legislators				
	1	2	3	4	5
1	167	95	88	74	137
2	95	167	130	120	110
3	88	130	167	105	97
4	74	120	105	167	85
5	137	110	97	85	167

task is to instruct the machine to record the number of times every man votes (has a 1 in his column) with each other man.

For both machines the cards must be sorted one column at a time. All cards which have a 1 in the sorted column are fed into the machine for tabulation of the *jo-matrix*. The matrix must be symmetrical. See Table 2. This is the check on the tabulation. Cards may be summary punched with the same totals that appear on the printed forms. These summary cards may be useful for further matrix manipulation or for larger summaries of the data, if these are part of the experimental design.

The 402 machine allows us to compare as many as 12 columns at a time. Each time a run is made, the control wire must be moved to the column on which the cards have been sorted. If  $n > 12$ , the wiring must be changed to pick up from the next set of 12 columns, 13-24, etc. We then begin sorting with column 1 and again run the entire gamut. Machines normally emit an impulse when two readings are unequal. The problem of wiring is to allow an impulse to be freed when two readings are equal, i.e., when there is a 1 in the sorted column and in any of the other columns being compared. This is accomplished by wiring from the comparing exit to the pilot selectors' digit pickup. The machine is wired for addition, minor programming, and printing of totals. If a percentage matrix is desired, the reciprocal of



the total number of votes is emitted into a counter entry. The number of significant digits required for accuracy must be borne in mind in computing this reciprocal.

The 101 machine compares as many as 60 columns at one time. The deck is sorted on one column and all cards with a 1 punch are run through the machine. The machine will print the total *jo*'s for 60 men with no wiring changes necessary. The essential wiring for the 101 machine provides that a 1 from the digit emitters be fed into the recode pickup, which has been wired so that the impulse passes from column to column. The recode selectors are also wired together. A wire runs from the count *to* to the recode selectors, and then from the unit counters to count return. At least one subtraction plug must be wired. The sort select switch is set at the 2 position. If the legislature exceeds 60 men, it is profitable to make enough decks to account for all of the men without rewiring the 101 machine. If there were 90 men, deck A could list men from districts 1-30 and 31-60; deck B from districts 1-30 and 61-90 (in columns 31-60); and deck C districts 31-60 (in the first 30 columns) and 61-90 (in the second set of 30 columns). The printed matrix is then spliced together. This method avoids the necessity for wiring changes and may also be employed with the 402 machine.

Since each investigator has his special problems of design, he will interpret these methods to suit himself. The 101 machine is the better one for even the smallest matrices. Fewer wiring problems are encountered, less time consumed, and the report is more readily checked. On the other hand, the 402 machine is more readily available at present installations.

### Discussion

The application of this method to psychological research may be made more explicit. This method may be applied to any dichotomous data. The vote is an excellent example. Sociometric choices provide a further major field of application.

Voting analyses and sociometrics have been criticized for failing to report reliabilities. We often study a handful of votes or adminis-

ter one sociometric and hope to describe or predict behavior. This tabulation technique makes possible the study of large numbers of votes and sociometric runs. Time or content matrices may be compared with their counterparts representing other time or content samples. To the degree that the *jo*'s are similar, we may speak of the *S*'s as being consistent and/or our measures as reliable. The application of this method to one legislature has been reported in the literature (12). We have since applied it to eight others. We found that behavior is significantly more consistent from issue to issue than from time to time. The research possibilities and practical applications are as broad as the ingenuity of the investigator.

We have alluded to the fruitfulness of having the data in matrix form. The reason for this is the development of matrix algebra and its application in factor analysis. As these statistical techniques become more refined, matrix data will assume greater importance. A few other possibilities are latent structure analysis (factor analysis applied to joint proportion) (13, 14), the difference method (19), matrix squaring or cubing (8), and the application of information theory (11). In addition to these reified techniques, clusters may be arbitrarily selected from the matrix without such refinement (3, 17, 21).

Finally, a major value of this method for applied studies is the speed with which the data may be tabulated. A legislature's votes on any day may be coded, punched, checked, and the *jo-matrix* tabulated overnight. Thus, a daily record may be kept of voting blocs. Weekly, monthly, or yearly summaries may be assembled. Matrices may also be tabulated according to special-interest legislation. In this manner, a legislator, citizenship committee, civic interest group, or social scientist could have at hand a daily, topic summary of the policy-body's voting patterns. A precise account of a group's sociometric development could similarly be made.

### Summary

A method for the quantitative treatment of dichotomous data is reported. This IBM method proceeds quickly from written rec-

ords to matrix form. Cards are mark-sensed with the data and punched and checked by machine. A matrix of joint-occurrences is tabulated by either of two IBM machines. The matrix has many practical applications. The method facilitates rapid, accurate analysis of political bodies, sociometrics, and other social data in dichotomous form.

Received September 5, 1953.

### References

1. Ash, P. The "liberalism" of Congressmen voting for and against the Taft-Hartley Act. *J. appl. Psychol.*, 1948, 32, 636-640.
2. Bales, R. F. *Interaction process analysis*. Cambridge, Mass.: Addison-Wesley, 1950.
3. Beyle, H. C. *The analysis of attribute-cluster-blocs*. Chicago: Univ. of Chicago Press, 1931.
4. Brimhall, D. R. and Otis, A. S. Consistency of voting behavior by our congressmen. *J. appl. Psychol.*, 1948, 32, 1-14.
5. Carlson, H. B. and Harrell, T. W. Voting groups among leading congressmen obtained by means of the inverted factor technique. *J. soc. Psychol.*, 1942, 16, 51-61.
6. Cohen, J. B. Note on Carlson and Harrell's *Factor analysis of voting among Congressmen*. *J. soc. Psychol.*, 1944, 20, 313-314.
7. Eberhart, J. C. Determinants of legislative behavior in the U. S. House of Representatives. *Psychol. Bull.*, 1942, 39, 595.
8. Festinger, L. The analysis of sociograms using matrix algebra. *Human Relat.*, 1949, 2, 153-158.
9. Fletcher, Mona. The use of mechanical equipment in legislative research. *Ann. Amer. Acad. polit. soc. Sci.*, 1938, 195, 168-175.
10. Gage, N. L. and Shimberg, B. Measuring senatorial "progressivism." *J. abnorm. soc. Psychol.*, 1949, 44, 112-117.
11. Garner, W. R. and Hake, H. W. The amount of information in absolute judgments. *Psychol. Rev.*, 1951, 58, 446-459.
12. Grace, H. A. A quantitative case study in policy science. *J. soc. Psychol.*, in press.
13. Green, B. F., Jr. A general solution for the latent class model of latent structure analysis. *Psychometrika*, 1951, 16, 151-166.
14. Green, B. F., Jr. Latent structure analysis and its relation to factor analysis. *J. Amer. Statist. Ass.*, 1952, 47, 71-76.
15. Harris, C. W. A factor analysis of selected senate roll calls, 80th Congress. *Educ. psychol. Measmt.*, 1948, 8, 582-591.
16. Keefe, W. J. Party government and lawmaking in Illinois General Assembly. *Northwestern Univ. Law Rev.*, 1952, 47, 55-71.
17. Klingberg, F. L. Studies in measurement of the relations among sovereign states. *Psychometrika*, 1941, 6, 335-352.
18. Lowell, A. L. The influence of party upon legislation in England and America. *Ann. Rep. Amer. Hist. Ass.*, 1901, 1, 321-542.
19. Osgood, C. E. and Suci, G. J. A measure of relation determined by both mean difference and profile information. *Psychol. Bull.*, 1952, 49, 251-262.
20. Rice, S. A. *Quantitative methods in politics*. New York: Knopf, 1928.
21. Tryon, R. C. Comparative cluster analysis. *Psychol. Bull.*, 1939, 36, 645-646.
22. Turner, J. Party and constituency: pressures on Congress. *Johns Hopkins Univer. Stud. in hist. polit. Sci.*, 1951, 69, No. 1.

## A Comparison of Two Methods of Measuring the Attention-Drawing Power of Magazine Advertisements

Joseph Tiffin and Darvin M. Winick

*Division of Applied Psychology, Purdue University*

In measuring the effectiveness of advertisements it is of primary importance to be able to measure the initial attention-drawing power or eye appeal of an advertisement. This fact is obvious, since an advertisement which is not seen cannot accomplish its intended purpose. One of the methods which has been used to accomplish this measurement of attention-drawing power is eye movement photography. This method produces an objective photographic record of the eye movements of a subject while he is observing one or several advertisements. If an experimental design permitted pairs of advertisements to be presented to the subject, the photographic record taken by an eye camera would indicate which of the two advertisements the subject preferred to observe. It is possible, then, by totaling the preferences of several subjects to scale a set of advertisements according to their relative attention-drawing power.

Although the eye movement photographic method produces an objective record of the subject's preferences, the method has several disadvantages:

1. The advertisements must be presented to one subject at a time and the presentation becomes relatively time consuming if large numbers of subjects are to be used.
2. The transportation and assembly of the necessary equipment is cumbersome.
3. The necessary task of frame by frame reading of the film record is laborious and time consuming.

It was the purpose of this study to investigate the possibility that a less time consuming method of measuring the attention-drawing power of advertisements will essentially scale advertisements in agreement with the scaling produced by eye movement photographic methods. Specifically, the investigation dealt with the relationship between scalings produced by a group tachistoscopic

method and scalings produced by the Purdue Eye Camera (2).

### Procedure

Ten advertisements to be scaled were selected from current issues of popular weekly magazines. All advertisements were in color and full page in size. The subjects for the study were 154 students in college psychology classes, education classes, and adult education classes.

*Tachistoscopic Presentation.* For the tachistoscopic part of the study the ten advertisements were reproduced on 35 millimeter colored transparencies which were individually mounted in  $1\frac{1}{4}'' \times 2''$  cardboard slide mountings. A special brass holder for pairs of these mounted transparencies was designed to fit into the slide carrier of a standard  $3\frac{1}{4}'' \times 4''$  lantern projector. It was possible, then, to project together on a screen any two of the ten advertisement slides. The brass holders could be slipped in and out of the slide carrier in the same manner as standard  $3\frac{1}{4}'' \times 4''$  lantern slides. In order to speed up the presentation two of the brass holders were constructed so that one pair of slides could be readied while another pair was in position to be projected on the screen.

A tachistoscopic shutter was mounted over the front lens of the projector so that the amount of time the advertisements were observed on the screen could be accurately controlled. In this study each possible pair of the ten advertisements was presented to the subjects for .5 seconds. Approximately 20 minutes were required for the complete set of 45 presentations.

Immediately following the presentation of each pair the subjects were asked to indicate on a prepared answer sheet which advertisement of the two they would look at if they were given a second look. The preferences of each subject for each advertisement were then determined from the answer sheets.

*Eye Movement Photography.* From the subjects participating in the tachistoscopic presentation, 36 were randomly selected to return for a second experiment in which the relative attention-drawing power of the ten advertisements was measured by use of the Purdue Eye Camera (2). This camera consists of a table stand on which two actual advertisements are placed, a half-



Table 1

Mean Number of Preferences for Random Halves of the Subjects Using a Group Tachistoscopic Presentation

Advertisement	Mean Number of Preferences for Halves of Subjects	
	Random Half A	Random Half B
1	7.10	7.09
2	6.54	6.64
3	5.70	5.22
4	5.26	5.43
5	3.82	4.23
6	3.92	3.93
7	3.66	3.78
8	3.21	3.17
9	3.13	3.05
10	2.65	2.45

silvered mirror placed directly in front of the stand, and an eight millimeter motion picture camera mounted in front of and above the mirror. A motor drive is used to keep the camera speed constant at 2.7 frames per second. The subject is seated in front of the stand and is able to view the advertisements through the half silvered mirror. The reflection on the mirror of the upper part of the subject's face is photographed by the motion picture camera.<sup>1</sup>

It is possible, by projecting the produced film a single frame at a time, to identify which advertisement the subject is looking at in each frame. A 35 millimeter strip film projector which had been converted for use with eight millimeter film was available for this single frame projection. A count of the number of frames during which the subject's eyes were fixed on a particular advertisement gives a measure of the amount of time in which the subject was looking at that advertisement. This amount of time spent on an advertisement was used as one measure of attention-drawing power. Another measure of attention-drawing power, the total first fixations on a particular advertisement, was also used. This was a measure of the number of times a particular advertisement was looked at first by the subjects during the paired presentation.

If each subject was to view all possible pairs of ten advertisements, it would be necessary to present 45 pairs and each advertisement would be presented nine times. In the eye camera experiment, however, it was felt that each subject should view each advertisement only once. In order to accomplish this only five pairs were pre-

sented to each subject. In this manner nine subjects were needed to complete each total pairing. The 36 subjects used in the eye camera experiment actually represented four complete pairings of the ten advertisements.

## Results

**Reliability.** From the results of the tachistoscopic presentation the preferences of all 154 subjects for each advertisement were totaled. The subjects were then randomly split into two groups and the product moment correlation (1) between the total preferences of these groups was computed and used as a measure of the reliability of the tachistoscopic method. The split-half correlation found was .98. When the Spearman-Brown formula (1) was applied to find an estimate of the expected correlation for double the number of judges, an  $r$  of .99 was found. In Table 1 the mean preferences for each half of the subjects are shown.

In order to investigate the reliability of the first eye camera measure, the first looks at each advertisement by random halves of the subjects were totaled. The  $r$  between these halves was .58. The Spearman-Brown formula established the  $r$  for double the number of judges to be .73. The mean number of first looks for random halves of the subjects are shown in Table 2. In the second eye camera method the first ten frames of film showing the subject looking at a pair of

Table 2

Mean Number of First Looks for Random Halves of the Subjects Using the Purdue Eye Camera

Advertisement	Mean Number of First Looks for Halves of Subjects	
	Random Half A	Random Half B
1	6.50	5.50
2	6.00	8.00
3	5.00	6.00
4	6.50	4.00
5	2.00	3.50
6	3.50	2.50
7	3.50	4.00
8	4.50	2.50
9	3.50	4.00
10	4.00	5.00

<sup>1</sup> A more detailed description of the camera can be found in an unpublished thesis by Karslake (3), and in an article by the same author (2).

advertisements were considered. The number of frames in which the subject looked at a particular advertisement was first determined, and from this the total number of frames in which random halves of the subjects looked at each advertisement was totaled. Since the eight millimeter camera was motor driven and its speed constant at 2.7 frames per second, it is possible to convert the total frames measure into total seconds looked at a particular advertisement. In Table 3 the mean number of seconds spent by random halves of the subjects on each advertisement is shown. The  $r$  between the two halves for this method was found to be .50. The Spearman-Brown estimate of reliability was .67.

*Comparison of the Different Methods.* In order to determine the relationship between the different methods of measuring the attention-drawing power of advertisements, product moment correlations were computed between the relative attention values found by the tachistoscopic presentation method and each of the two eye camera measures. In Table 4 the mean number of tachistoscopic preferences, first looks, and seconds spent on each advertisement are shown for all subjects. The correlation between the results of the tachistoscopic and eye camera, first look, methods was found to be .79. A correlation of .83 was found between the results of the

Table 3

Mean Number of Seconds Spent by Random Halves of the Subjects Using the Purdue Eye Camera

Advertisement	Mean Time in Seconds Spent by Halves of Subjects	
	Random Half A	Random Half B
1	21.3	18.7
2	21.8	18.3
3	18.9	17.8
4	15.0	21.3
5	13.1	13.0
6	15.6	17.4
7	15.4	13.7
8	13.3	15.0
9	16.9	15.6
10	15.4	15.9

Table 4

Mean Number of Tachistoscopic Preferences, First Looks, and Seconds Spent for All Subjects

Advertisement	Mean Number of Tach. Preferences	Mean Number of First Looks	Mean Number of Seconds Spent
1	7.10	6.00	20.0
2	6.59	7.00	20.1
3	5.46	5.50	18.3
4	5.34	5.25	18.1
5	4.03	2.75	13.0
6	3.93	3.00	16.5
7	3.72	3.75	14.5
8	3.19	3.50	14.1
9	3.09	3.75	16.2
10	2.55	4.50	15.6

tachistoscopic preferences and the number-of-seconds-spent measure of the eye camera presentation.

Since the reliability of the two eye camera measures was lower than the reliability of the tachistoscopic measure, probably due primarily to the small number of complete pairings, it would be of interest to know what the correlations between the tachistoscopic results and each of the two eye camera measures would be if the latter were perfectly reliable. These correlations can be estimated by correcting for attenuation due to the imperfect reliability of the eye camera (criterion) measures (1, p. 530). These corrected correlations were .86 between the tachistoscopic measures and the eye camera, first look, measures; and .99<sup>2</sup> between the tachistoscopic results and the eye camera "number of seconds spent" measures.

Summary and Conclusions

In this investigation it was found that the attention-drawing power of advertisements can be scaled by the paired-comparison tachistoscopic method with a reliability of .99. Correlations of .86 and .99 were found between the tachistoscopic method and two eye camera measures. These  $r$ 's were the result of correcting the obtained correlations for the unreliability of the eye camera measures.

<sup>2</sup> Actual arithmetic results in an  $r$  of 1.03. For a discussion of  $r$ 's greater than unity, see McNemar (4, p. 136).

The relationships indicate that the group tachistoscopic method as used in this study will scale advertisements in essentially the same order as eye camera methods when attention-drawing power is considered. This fact is important to people interested in advertising research for several reasons:

1. The tachistoscopic method lends itself easily to group presentation and enables large numbers of subjects to be reached.

2. A standard, easily transportable, slide projector is the only equipment needed to make the tachistoscopic presentation.

3. Preferences on prepared answer sheets may be quickly totaled by hand or machine methods.

In situations where eye movement photography could be used to measure the attention-drawing power of advertisements, the results

of this study indicate that a considerable saving of time and energy can be effected by use of a group tachistoscopic presentation.

Received June 19, 1953.

#### References

1. Guilford, J. P. *Fundamental statistics in psychology and education*. (2nd Ed.) New York: McGraw-Hill, 1950.
2. Karslake, J. S. The Purdue Eye-Camera: A practical apparatus for studying the attention value of advertisements. *J. appl. Psychol.*, 1940, 24, 417-440.
3. Karslake, J. S. *A simple and direct method for investigation of the attention value of advertising copy through eye movement photography*. An unpublished thesis, Purdue University, 1939.
4. McNemar, Q. *Psychological statistics*. New York: Wiley & Sons, 1949.



## Applied Psychology in Action

### Legal Status of Advertising and Marketing Psychology Experts

An important U. S. District Court decision by Judge Robert C. Bell, District of Minnesota, was handed down on September 4, 1953 at St. Paul, Minn. The Court admitted the testimony of two experts in the field of advertising and marketing psychology. These experts had been engaged by the U. S. Food and Drug Administration to interpret advertising copy and to determine its impact on a sample of 200 prospective purchasers of a drug. As a result of the success attained it is probable that advertising and marketing psychologists will be increasingly used in prosecutions involving the fraudulent and misleading use of labels and advertisements in the marketing of drugs and foods.

The case revolved around a full page newspaper advertisement of "Tryptacin." The label on the drug itself did not contain directions for use in the treatment of stomach ulcers although this is a condition for which the drug is intended and for which the drug is suggested and recommended in its advertising.

The defense contended that the advertisement represented only that "Tryptacin" is intended for use as an antacid or a palliative for acid pain. The defense evidence consisted of the testimony of two representatives of a firm which handles "Tryptacin" advertising and the testimony of a number of physicians. The two advertising men testified that, in their opinion, the advertisement offered "Tryptacin" as a means of relieving acid pain and not of curing stomach ulcers. They also testified that they had shown the advertisement to a number of their associates in the advertising business, to newspaper censorship boards, and to other persons and not a single person received the impression that the advertisement offered a cure for stomach ulcers. The physicians who testified for the defense stated that they had discussed the advertisement with doctors, nurses, patients, and other persons and again no one got the idea that the product would cure stomach ulcers. The Court noted that these witnesses

did not offer any written evidence concerning their interviews. Furthermore, it did not appear to the Court that the interviews were systematically conducted. The Court went on to state: "The likelihood of error or prejudice developing in the course of such interviews would seem to be great, particularly since *none of the witnesses of claimant, including both advertising men and doctors, were qualified by education or experience in the taking of formal public opinion surveys.*" (Italics added)

Judge Bell based his decision on: (1) reading and examining the advertisement; (2) hearing the testimony of two experts in the field of advertising and marketing psychology (Howard P. Longstaff of the University of Minnesota and James N. Mosel of George Washington University); (3) the testimony of two persons who purchased the drug in the belief that the advertisement offered a cure for stomach ulcers; and (4) the testimony of a specialist in internal medicine who has treated many cases of stomach ulcers and who testified that in his opinion the ulcer patient would get the impression from the advertisement that the drug was offered as a cure for stomach ulcers.

The Court commented on the testimony of Longstaff and Mosel as follows: "they presented exhaustive analyses of the content of the advertisement and the effect which it was intended to have upon the prospective purchaser of the drug. Such testimony is admissible to determine the meaning of an advertisement. *Federal Trade Commission v. National Health Aids, Inc.*, 108 F. Supp. 340 (D. Md.).

"Moreover, Dr. Mosel introduced evidence relative to two hundred individuals whom he surveyed concerning the impression which they received from the 'Tryptacin' advertisement. A substantial portion of those interviewed indicated that they received the impression from the advertisement that 'Tryptacin' would 'stop,' 'cure' or otherwise bring

about some permanent relief of ulcers. The forms filled out by the individuals questioned, interview cards, and tabulations made by Dr. Mosel of the answers received, were placed in evidence."

The Court thereupon upheld the seizure of 363 cases, more or less, of the drug and

assessed the costs of the judicial proceedings against the defense.—Source: Letter dated September 23, 1953 from Division of Regulator Management of the Food and Drug Administration together with enclosures consisting of Findings of Fact and Conclusions of Law and Memorandum Opinion.

THE JOURNAL OF APPLIED PSYCHOLOGY  
Vol. 38, No. 4, 1954

## Reporting Employment Test Scores to Supervisors \*

Clifford E. Jurgensen

*Ass't Vice President—Personnel, Minneapolis Gas Company*

One of the persistent problems in the field of Industrial Psychology is that of reporting employment test scores to persons untrained in the field of tests and measurements. Such persons include supervisors, top management, and on occasion, perhaps, the applicant himself. It is simple enough to advise that test scores should not be given persons untrained in test interpretation. In actual practice, however, such advice must often be ignored.

Training in test interpretation can and should be given insofar as is possible. However, such training cannot possibly reach all persons involved. Further, it is unlikely that training can be sufficiently intensive and extensive to train adequately any of the persons involved. Therefore, it is necessary and desirable to simplify test score interpretation to the greatest possible extent.

The procedure discussed here consists of a profile chart on which percentile scores are plotted on a linear continuum. The chart, shown in Figure 1, is based on normal probability tables in which percentile ranks are plotted in accordance with  $z$ -score units. These units effectively overcome the difficulty presented by the fact that the difference between the 90th and 99th percentiles is not equivalent to the difference between the 40th and 49th percentiles.

Although carefully designed experiments with adequate controls are lacking, experience indicates that lay people tend automatically to make reasonably correct interpretation of scores inasmuch as they are likely to interpret scores on the basis of where the  $X$ 's appear on the profile. For example, it is not uncommon to hear remarks such as "His score on mechanical reasoning is about half way between his highest and lowest scores." Such remarks are based on profile plotting (and therefore  $z$ -scores) and do not correspond to an average percentile rank.

Although it has been found that lay people typically interpret scores graphically, and therefore linearly insofar as standard scores are concerned, the profile does give two verbal interpretations to facilitate communication or record purposes. One of these is the well known percentile rank which is labeled on the profile as "per cent of group having lower score." The other is a general verbal interpretation of the score in terms of commonly used adjectives. A column headed "Test or Measurement" is used to give the type of test in functional terms rather than the name of the specific test. A column headed "Basis of Comparison" is used to give the norm group on which test scores are profiled.

The profile chart mentioned above is a simplification of a similar chart used within the Personnel Department with persons trained in test interpretation. This original chart permitted interpretation on four, rather

\* This material contains the gist of a part of the presentation by Jurgensen in a panel discussion on "Philosophy of Testing" before the Minneapolis Vocational Guidance Association on April 29, 1954.

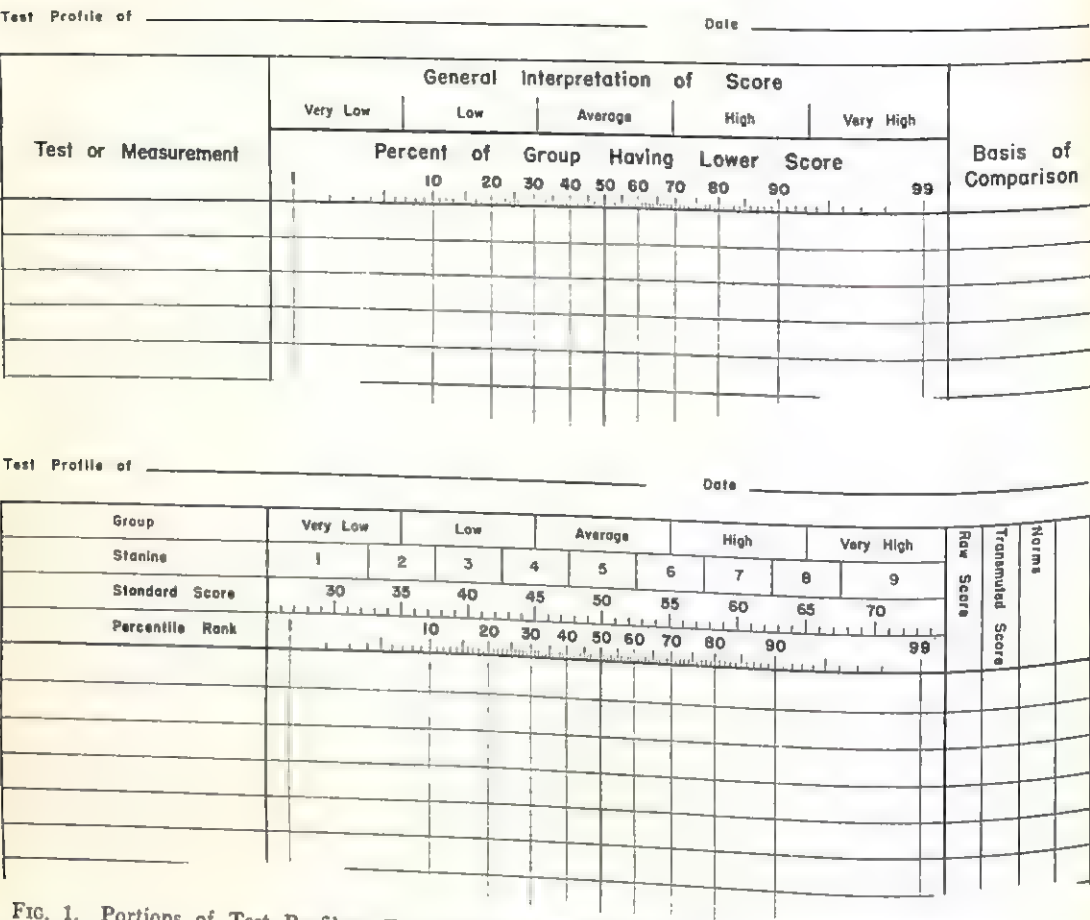


FIG. 1. Portions of Test Profiles. Test Profile Report Form at the top is non-technical for use with supervisors, applicants, etc. The Report Form at the bottom is for technical uses within the Personnel Department.

than two bases. These are: group description, stanine, standard score, and percentile rank. Instead of a single column labeled "Basis of Comparison," the original chart contained four columns. These consisted of raw score, transmuted score (percentile rank, standard score, stanine, or other such score), norm group, and a fourth column could be used for any additional data desired. This original, and more complicated, chart contains the same advantages as the simplified version insofar as interpretation based on location of X. However, although terms such as stanine and standard score do not affect score interpretation, lay people feel uneasy about a chart which they do not fully comprehend. The simplified profile has therefore been found to contain all of the advantages without containing the disadvantages of the original chart.



## Book Reviews

Marketing and Social Research Division of the Psychological Corporation. *The measured effectiveness of employee publications*. New York: Association of National Advertisers, Inc., 1953. Pp. 109. \$10.00.

Here is a well designed study, the results of which are reported in a beautiful, lithographed brochure replete with illustrations, simple tables of results expressed as percentages, a minimum of verbiage, and a maximum of white spaced margins. The overall page size is 14 inches by 11 inches. Presumably this is the kind of expensive and expensive-looking report that consultants and consulting organizations *believe* will be read by top brass in business and industry. It is in decided contrast to the form of report adopted by scholars in reporting the detailed results of quantitative studies in the scientific journals and monographs. The very form of this report, in this reviewer's opinion, poses a practical psychological problem: do high level executives really prefer this type of advertising lay-out report? Is the stereotype of the "busy executive" correct which assumes that what is put before him for his serious attention and study must be presented in a form so that "he who runs will read"?

But let's get on with a review of the contents of this report. As the sub-title states, it is a study of readership, penetration, and readability of seven employee publications. The sponsor was the ANA Public Relations Committee and the study itself was made by the Psychological Corporation with Charles L. Vaughn serving as technical director. It was aimed at finding out what employees will read and believe.

A foreword, an introduction dealing with the role of employee publications in the total area of business communications, the objectives of the study, the methods used in the investigation, and a general summary of results are then followed by a pictorial, tabular, and verbal description of the detailed results.

In general, the results show the employee publication to be one of the best of available sources of information about the company—

better than such sources as the first-line supervisor, the union steward, and meetings. An incredible 97 per cent of the 1,800 in-plant interviews indicate belief in what they read in the company magazine or newspaper. Readership was likewise quite high, namely 90 per cent reported they had read at least one of the two most recent issues and, on the average, 78 per cent reported that they read the publications regularly. Thoroughness of readership, however, was much less.

The industrial psychologist and the applied social psychologist will be especially interested in the reported relationships between "leftist" and "rightist" attitudes of employees interviewed and the extent of their readership and in the Flesch readability scores of these seven publications. In regard to the latter, as is usual, the publications are written at a level of difficulty that is too high for over one-third of the rank-and-file employees. Of more importance, little relationship between readership and readability was found. This finding, however, may be regarded as throwing doubt on the method of measuring readership which was used rather than as evidence tending to discredit the importance of simplified language in reaching employees with limited amounts of education.

The reviewer has little to criticize with respect to what is actually presented in this report. Furthermore, there are many commendable features such as the frankness of the plant-by-plant comparisons, the wealth of pictorial illustrations of "good" and "poor" features of these company publications, and the reproduction, in the appendix, of the interview schedule used to measure readership and attitudes. It is obvious that high-level professional competence is reflected. However, one misses any reference to other relevant studies to which the scholarly business executives could turn for further information if he so desired. The reviewer suspects there are many more studious business executives than the advertising fraternity responsible for the form of the present report

would believe possible. Finally, it would have been of value to the serious student of industrial communications to give the statistical constants such as means and standard deviations, and coefficients of correlation in an appendix so that findings in the present study might be compared with reports of similar scientific studies.

Donald G. Paterson

*The University of Minnesota*

### Comment on Preceding Review

Paterson's remarks in regard to the elaborateness of the presentation accentuate the rather interesting differences in frames of reference. After considerable discussion, we decided to make the publication simple because we felt that executives were tiring of the glossy four-color jobs! Actually, only the general summary and preceding page were thought to be of interest to top management, and these two pages will probably be printed separately for that group.

We were rather concerned with the low and occasionally negative correlations between Flesch scores and readership. The explanation, I believe, is to be found more in the nature of the articles than in the weaknesses of the formula or the measures of readership, although neither is perfect by any means.

What happens, I suspect, is that the timely and important material may often be hastily written at the last minute, and may come from persons not concerned with writing readably. "Dravo Bids," a feature illustrated in our publication, actually tells the workers indirectly whether they are going to have jobs or not, yet it abounds with long words and big figures. I can verify from intensive interviewing of my own that even the poorly educated "sand hogs" literally pore through it. There are other similar features.

The travesty is that compelling material written at a very difficult level may lead the inept reader to some rather bizarre conclusions indeed.

Charles L. Vaughn

*The Psychological Corporation*

Anon. *Army personnel tests and measurements*. TM12-260, Department of the Army. Washington, D. C.: U. S. Government Printing Office, 1953. Pp. 125. \$.55.

This is a good little summary of the use of tests and rating procedures in the Army. It reads much like a standard text on employment psychology condensed and written down to the level of readers without a psychology background. For psychologists in the Army it might serve as a useful refresher and almost approximates a manual. For other Army personnel needing some familiarity with the field, it would be very helpful if read carefully and, preferably, with an elementary statistics text on the side. The monograph covers test construction, criteria, scoring of tests (especially standard scores), reliability and validity, the use of profiles in classification, achievement tests, self-description and rating scales (including forced choice), test administration and scoring.

The work has a number of commendable features. It is concise and there is not a word wasted. Effective use is made of graphic materials—some of them quite ingenious. There is interesting adaptation of military terminology to conventional psychological presentation. For instance, reliability and validity are interpreted in terms of "calculated risks." The treatment is down to earth and practical, but entirely scientific withal.

There is always the problem of how to handle statistics in a work like this. The present authors employ conventional statistical terminology, but do not indicate how anything is computed. There is a frequent suggestion that "any statistics book" covers some particular item. The authors do about as well as could be done under the circumstances with brief explanations of some statistical notions and graphic materials to clarify the explanation. According to an insert the major responsibility of the work appears to have been carried by Baier, Bayroff, and Rundquist. They are to be congratulated on having done an interesting and useful minor piece of work.

Harold E. Burt

*The Ohio State University*

Buros, Oscar K., editor. *The fourth mental measurements yearbook*. Highland Park, N. J.: The Gryphon Press, 1953. Pp. xxiv + 1163. \$18.00.

Most reviews of earlier editions of the *Mental Measurements Yearbook* have begun with accolades. The reviewer of this latest edition sees no reason to deviate from this course: Buros' Fourth Mental Measurements Yearbook is a monumental work, even longer than the previous edition and of inestimable value to purveyors and users of information about tests. "The (825-page) section 'Tests and Reviews' lists 793 tests, 596 original reviews by 308 reviewers, 53 excerpts from test reviews in 15 journals, and 4,417 references on the construction, validation, use, and limitations of specific tests. . . . The (267-page) section 'Books and Reviews' lists 429 books on measurement and closely related fields and 758 excerpts from book reviews in 121 journals." The series of detailed indexes remains an excellent feature of the volume.

Projective tests, aptitude test batteries, and tests for specific vocations all receive noticeably more attention in the present volume than in the Third Yearbook. Some 19 projective tests are mentioned for the first time in the yearbook series and 631 new journal references on the Rorschach (one-seventh of the total number of journal references on tests) bring the total in the yearbook series to an impressive 1,217. The one page devoted to three aptitude test batteries in the Third Yearbook has become 37 pages devoted to nine such batteries in the current work. That the aptitude test battery is a relatively recent development is made clear by the post-World War II dates on seven of the nine batteries. That 45 tests for specific vocations are listed

in the current yearbook, as opposed to 10 in the Third Yearbook, is partly the result of the only recently won permission to review several of these tests, partly a reflection of the continued efforts of professional schools to improve selection procedures.

Past reviewers have argued for changes in editorial policy, notably for the exclusion of tests which do not meet certain predetermined criteria. The present reviewer chooses to concern himself with only one aspect of editorial policy: the exclusion of tests thoroughly reviewed in previous yearbooks for which there has been no new edition since the last yearbook.

Unless it can be assumed that all yearbook users know they must also consult previous editions, they may not become aware of the existence of some established tests. At least a half dozen of the best known, most used (and frequently most carefully studied) tests of manual dexterity are not mentioned in the Fourth Yearbook, nor is the well-known Minnesota Clerical Test. Current yearbooks should at least list tests previously reviewed with a cross reference to the appropriate volume. Exclusion criteria might be developed so that such lists would not be cluttered with the measurement whims of the century.

Added features require space, and space has always been a problem for Buros. The "Books and Reviews" section appears to offer less that is new and to serve a more limited readership. To the extent that this section is a drain on the "Tests and Reviews" section, it is here that space economies should be effected.

Charles N. Morris

*Teachers College, Columbia University*



## New Books, Monographs, and Pamphlets

Books, monographs, and pamphlets for listing and possible review should be sent to Donald G. Paterson, Editor, Department of Psychology, University of Minnesota, Minneapolis 14, Minnesota.

- Problems of consciousness.* Harold A. Abramson, Editor. New York: The Josiah Macy, Jr. Foundation, 1954. Pp. 177. \$3.25.
- Rorschach responses in the aged.* Louise Bates Ames, Janet Learned, Ruth W. Metraux, and Richard N. Walker. New York: Paul B. Hoeber, Inc., Medical Book Department of Harper & Brothers, 1954. Pp. 244. \$6.75.
- Psychological testing.* Anne Anastasi. New York: The Macmillan Company, 1954. Pp. 240. \$4.25.
- The exteriorization of the mental body.* James Baker, Jr. New York: The William-Frederick Press, 1954. Pp. 32. \$1.50.
- Psychology of personnel in business and industry.* Second Edition. Roger M. Bellows. New York: Prentice-Hall, Inc., 1954. Pp. 467. \$7.35.
- Employment psychology: the interview.* Roger M. Bellows and M. Frances Estep. New York: Rinehart & Company, Inc., 1954. Pp. 295. \$4.25.
- After high school—what?* Ralph F. Berdie. Minneapolis: University of Minnesota Press, 1954. Pp. 240. \$4.25.
- Columbia mental maturity scale.* Bessie B. Burgemeister, Lucille Hollander Blum, and Irving Lorge. Yonkers-on-Hudson, N. Y.: World Book Company, 1954. Examiner's Kit: 100 items, and a comprehensive Manual. \$35.00. Individual Record Blanks are priced at \$.85 per package of 35.
- The sociology of work.* Theodore Caplow. Minneapolis: University of Minnesota Press, 1954. Pp. 330. \$5.00.
- Manual of child psychology.* Second Edition. Leonard Carmichael, Editor. New York: John Wiley & Sons, Inc., 1954. Pp. 1,295. \$12.00.
- Sociology perspective.* Ely Chinoy. New York: Doubleday and Company, Inc., 1954. Pp. 58. \$.85.
- Introduction to logic.* Irving M. Copi. New York: The Macmillan Company, 1953. Pp. 472. \$4.00.
- Symbolic logic.* Irving M. Copi. New York: The Macmillan Company, 1953. Pp. 472. \$5.00.
- Religion and human behavior.* Simon Doniger, Editor. New York: Association Press, 1954. Pp. 233. \$3.00.
- Production guides and controls for the modern executive.* M. J. Dooher, Editor. New York: American Management Association, 1953. Pp. 52. \$1.25.
- Stepping up office efficiency.* M. J. Dooher, Editor. New York: American Management Association, 1953. Pp. 46. \$1.25.
- Streamlining office equipment and service.* M. J. Dooher, Editor. New York: American Management Association, 1953. Pp. 35. \$1.25.
- Gearing up for better production.* M. J. Dooher, Editor. New York: American Management Association, 1953. Pp. 58. \$1.25.
- The human side of the office manager's job.* M. J. Dooher, Editor. New York: American Management Association, 1953. Pp. 40. \$1.25.
- A critical look at the insurance buyer's role.* M. J. Dooher, Editor. New York: American Management Association, 1953. Pp. 35. \$1.25.
- Maintaining a dynamic insurance program.* M. J. Dooher, Editor. New York: American Management Association, 1953. Pp. 44. \$1.25.
- Industry at the bargaining table.* M. J. Dooher, Editor. New York: American Management Association, 1954. Pp. 51. \$1.25.
- Selling costs and market potential: controls and guides.* M. J. Dooher, Editor. New York: American Management Association, 1954. Pp. 38. \$1.25.
- Modern learning theory.* William K. Estes, Sigmund Koch, Kenneth MacCorquodale, Paul E. Meehl, Conrad G. Mueller, William N. Schoenfeld, and William S. Verplanck. New York: Appleton-Century-Crofts, Inc., 1954. Pp. 424.

- Mind and performance.* Harold Kenneth Fink. New York: Vantage Press, 1954. Pp. 113. \$3.00.
- Human behavior in industry.* William W. Finlay, A. Q. Sartain, and Willis M. Tate. New York: McGraw-Hill Book Company, Inc., 1954. Pp. 247. \$4.00.
- A psychological glossary.* D. C. Fraser. Cambridge, England: W. Heffer & Sons, Ltd., 1954. Pp. 40. 3s. 6d. net.
- Methods of research.* Carter V. Good and Douglas E. Scates. New York: Appleton-Century-Crofts, Inc., 1954. Pp. 896. \$5.50.
- The life and ideas of the Marquis De Sade.* Geoffrey Gorer. New York: The British Book Centre, Inc., 1954. Pp. 244. \$3.50.
- Child psychology.* Fourth Edition. Arthur T. Jersild. New York: Prentice-Hall, Inc., 1954. Pp. 676. \$6.00.
- The practice of psychotherapy.* C. G. Jung. New York: Bollingen Series, 140 East 62nd Street, 1954. Pp. 377. \$4.50.
- Know your reader.* George R. Klare and Byron Buck. New York: Hermitage House, Inc., 1954. Pp. 192. \$2.95.
- The technique of handling people.* Revised Edition. Donald A. and Eleanor C. Laird. New York: McGraw-Hill Book Company, Inc., 1954. Pp. 189. \$3.75.
- Towards an understanding of juvenile delinquency.* Bernard Lander. New York: Columbia University Press, 1954. Pp. 143. \$3.00.
- Your child and his art.* Viktor Lowenfeld. New York: Macmillan Company, 1954. Pp. 186. \$6.50.
- Break down the walls.* John Bartlow Martin. New York: Ballantine Books, 1954. Pp. 310. Paperbound edition \$5.00. Hardbound edition \$3.50.
- A new approach to office management: integrated data processing through common language machines.* Elizabeth Marting, Editor. New York: American Management Association, 1954. Pp. 62. \$2.50.
- People's Padre.* Emmett McLoughlin. Boston: Beacon Press, 1954. Pp. 288. \$3.95.
- How to enjoy yourself.* Albert A. Ostrow. New York: E. P. Dutton & Co., Inc., 1954. Pp. 259. \$2.95.
- Psychology.* William J. Pitt and Jacob A. Goldberg. New York: McGraw-Hill Book Company, Inc., 1954. Pp. 414. \$4.50.
- Psychology and life.* Fourth Edition. Floyd L. Ruch. New York: Scott Foresman and Company, 1954. Pp. 496. \$5.00.
- Letters to my daughter.* Dagobert D. Runes. New York: Philosophical Library, 1954. Pp. 131. \$2.50.
- Principles of industrial psychology.* Thomas Arthur Ryan and Patricia Cain Smith. New York: Ronald Press Company, 1954. Pp. 534. \$5.50.
- Selected writings of De Sade.* Leonard de Saint-Yves. New York: The British Book Centre, Inc., 1954. Pp. 306. \$6.75.
- Case studies in management development: theory and practice in ten selected companies.* Robert G. Simpson. New York: American Management Association, 1953. Pp. 140. \$2.50.
- A survey of management development: the quantitative aspects.* Joseph M. Trickett. New York: American Management Association, 1953. Pp. 64. \$1.25.
- Management education in American business.* Lyndall F. Urwick. New York: American Management Association, 1953. Pp. 136. \$1.50.
- An annotated bibliography of word association references important to marketing researchers.* James M. Vicary. New York: James M. Vicary Company, 20 East 60th Street. Pp. 5. Gratis.
- The education of employees: a status report.* Douglas Williams and Stanley Peterfreund. New York: American Management Association, 1953. Pp. 65. \$1.25.
- Personality through perception: an experimental and clinical study.* H. A. Witkin, H. B. Lewis, M. Hertzman, K. Machover, P. Bretnall Meissner, and S. Wapner. New York: Harper & Brothers, 1954. Pp. 571. \$7.50.
- Audio-visual materials: their nature and use.* Walter Arno Wittich and Charles F. Schuller. New York: Harper & Brothers, 1953. Pp. 564. \$6.00.
- Psychology in the nursery school.* Nelly Wolffheim. New York: Philosophical Library, 1953. Pp. 144. \$3.75.

- Journal of counseling psychology.* C. Gilbert Wrenn, Editor. Business Office: Room 2, Old Armory, Ohio State University, Columbus 10, Ohio. \$6.00 per year. \$1.75 per issue. Issued bi-monthly.
- Reading rapidly and well.* Revised Edition. C. Gilbert Wrenn and Luella Cole. Stanford, Calif.: Stanford University Press, 1954. Pp. 16. \$.15.
- The language of dynamic psychology.* Joseph W. Wulfeck and Edward M. Bennett. New York: McGraw-Hill Book Company, Inc., 1954. Pp. 111. \$4.00.
- Administration and the teacher.* William A. Yeager. New York: Harper & Brothers, 1954. Pp. 577. \$4.50.
- The pre-adolescent exceptional child.* Child Research Clinic of the Woods Schools. Langhorne, Pa.: The Child Research Clinic of the Woods Schools, 1953. Pp. 70. Gratis.
- This we believe about education.* Educational Advisory Committee and Council. New York: National Association of Manufacturers. Pp. 32.
- Studies in schizophrenia.* Tulane Department of Psychiatry and Neurology. Cambridge, Mass.: Published for the Commonwealth Fund by the Harvard University Press, 1954. Pp. 619. \$8.50.
- Statistics of public secondary day schools, 1951-1952.* U. S. Department of Health, Education, and Welfare. Washington 25, D. C.: Superintendent of Documents, U. S. Government Printing Office, 1954. Pp. 81. \$.35.



# Journal of Applied Psychology

VOL. 38, No. 5

OCTOBER, 1954

## Studies in Industrial Empathy: III. A Study of Supervisory Empathy in the Textile Industry \*

Wendell M. Patton, Jr.

Bruce Payne & Associates, Inc.

### Summary

The increasing difficulties and growing responsibilities of the position of supervisor, particularly in the realm of human relations suggest the need for the investigation of the ability of supervisors to understand both their superiors and subordinates. With this in mind, a study was undertaken of the empathetic ability of these supervisors and the extent to which this ability was related to other psychological variables.

Data were obtained from a large textile manufacturing plant producing prints and materials from spun rayon. The results are based on the replies of 54 secondhands or front-line supervisors, 18 members of top management, and a random sample of 243 out of 2,496 employees.

It was found that the secondhands were not empathizing effectively with either labor or management. Instead they were projecting, positively toward labor and negatively toward management. A social-psychological gap existed between labor and management which the supervisors were unable to perceive. Intelligence, education, and scores on the test, *How Supervise?* (5) were positively related to the supervisors' empathetic ability for both labor and management; age and supervisory experience were negatively related, while the particular shift and department in which a supervisor was employed apparently had no effect on empathetic

ability. Empathetic ability was no greater among supervisors who were considered by management to be the best than those considered by management to be the worst. Intercorrelations between related variables and the supervisors' empathy scores showed that the supervisors' own attitudes and knowledge were the chief factors influencing projection.

The findings indicated important individual differences in empathetic ability and the possibility of predicting from a regression equation the supervisor's ability to empathize with either labor or management.

### The Problem

Today the American industrial system has become a house divided against itself. In industrial enterprise the supervisor is the direct connecting link between labor and management. The increasing difficulties, complications, and responsibilities of textile supervision have made it increasingly necessary to devote more effort to determining some of the psychological characteristics of good leadership and of the men now occupying these positions. In the final analysis it is the supervisor who determines whether or not a given worker will keep his job or be fired or promoted. It is this supervisor who gives the orders and carries the directives of management to the workers. It is this same supervisor who has the only direct personal contact with the workers, and to these workers his actions and decisions are direct expressions of company policy. Since efficient supervision demands a two-way channel of communication, it appeared likely that those individuals who have the ability to "put

\* This paper is based upon the writer's doctoral research directed by Dr. H. H. Remmers, Purdue University. The dissertation, *A Study of Certain Psychological Variables Related to Supervision in the Textile Industry*, is on file in the Purdue University Library.

themselves in the other fellow's shoes" and anticipate their responses would best be able to carry the directives of management to labor and the needs and attitudes of labor to management. This ability (empathy) and its relation to other psychological variables of supervision constitute the basis of this study.

Background

Though the concept of empathy is of comparatively recent origin, the possibilities of its value in various situations has not escaped the attention of investigators. Remmers (8), for example, used this concept when he was called upon to develop an experimental design to test the procedures used to reduce the social-psychological gap between labor and management in a large industrial organization. Davidoff (1) was concerned with the reciprocity of empathy between Negroes and whites while Miller and Remmers (7) studied the psychological distance between organized labor and management. Interest in the attitudes of labor leaders toward industrial supervision was shown by Remmers and Remmers (9) while Richards (10) measured the empathetic ability of both labor and management. Travers (12), in a study of predicting public opinion, found that individuals tend to overestimate the percentage of the group being judged who feel as they themselves feel. Projection of this nature which renders the empathetic process difficult is readily observable about us in relationships such as between parents and children, teachers and pupils, and others.

The present study was attempted in part as a service project for the plant in which it was conducted and consequently leaves many facets untouched. Even so, many avenues for additional research were uncovered. The development of some objective measure of the concept of empathy as well as the study of the effect of training upon empathy would certainly be well worth while. If individuals can be taught to empathize more closely, the effect would far exceed the confines of an industrial situation. If this ability is not affected by training, then the problem becomes one of selection. The reciprocity of empathy would also be an interesting area of study. It has been suggested that empathetic ability is reciprocal in nature so that it is easier to empathize with an individual who has high empathetic ability than one who has low ability (4). A knowledge of the many variables affecting empathy and the empathetic ability within an individual at different times would also add to our meager information on this subject.

The results of these studies as well as Libo's (6) and Dymond's (2, 3, 4) suggest that empathetic ability is important for directing the work of others.

Procedure

Remmers (8) operationally defines empathy as "... having the subject or subjects predict the ordinal or cardinal position of another individual or group on one or more scales of defined psychological dimensions." The scale chosen for this study was *How Supervise?* (5). This test was administered to all front-line supervisors (secondhands), general foremen (overseers), members of top management and to a random sample of 10% of the employee group. It was also administered to the secondhands on two other occasions: once with instructions to answer each question as they believed management would answer it, and again with instructions to answer each question as they believed the employees would answer it. For the purpose of this study the scoring consisted of counting the correct responses as determined by the answer key. The index of empathy was computed by determining the difference between the predicted scores for a given group and that same group's actual mean score.

These same supervisors were also administered *The Adaptability Test* (11) which was designed to yield a general measure of intelligence. Information such as age, sex, experience and education was obtained from the personnel records and a personal history blank which all supervisors completed. Since no suitable production records were available for a criterion of supervisory efficiency, ratings of supervisors by superiors were used. Each supervisor was rated by at least three superiors and from these ratings a rank-order list was formed. Data from these sources served to test relevant hypotheses.

Results

The extent to which textile supervisors, labor and management understand the psychologically best methods of supervision is shown in Table 1. The front-line supervisors scored higher than labor but management scored higher than either the super-

Table 1  
Comparison of the Mean Scores of Labor, Front-line Supervisors and Management on *How Supervise?*

Group	Number	Mean	Standard Deviation	Standard Error of Mean
Labor	243	44.1	10.3	.66
Front-line Supervisors	54	48.1	8.5	1.17
Management	18	53.8	4.7	1.14



Table 2

Comparison Between the Actual Social-Psychological Distance Between Management and Labor as Measured by *How Supervise?* and the Front-line Supervisors' Prediction of this Distance

	Management	Labor	Difference
Actual Mean	53.84	44.14	9.70
Standard Error	1.14	.66	1.31
Predicted Mean	46.87	48.72	-1.85
Standard Error	1.16	1.04	1.55
Difference	-6.97	+4.58	11.55
Standard Error	1.62	1.23	2.06

visors or the workers. These differences were significant at the 1% level of confidence.

Table 2 shows a comparison between the actual social-psychological distance between management and labor and the supervisors' perception of this difference. The supervisors tended to overestimate the workers' knowledge of the best methods of supervision by a mean difference of 4.58. They also tended to underestimate management's knowledge of the best methods of supervision by a mean difference of 6.97. Both of these differences were significant at the 1% level of confidence. The social-psychological distance is indicated by the difference of 11.55 which is also significant at the 1% level of confidence. From the data shown in Table 2 it becomes obvious that the supervisors are not empathizing optimally with either management or labor. Instead they are projecting in both cases: positively toward labor and negatively toward management.

A Pearson  $r$  of  $+ .61$  was found between the supervisors' own scores on the test *How Supervise?* and the scores they predicted for

labor. A Pearson  $r$  of  $+ .44$  was found between the supervisors' own scores and the scores they predicted for management. Both correlations were significant at the 1% level of confidence. This relationship very clearly implies that one important reason for the failure of supervisors to empathize is the projection of their own attitudes and knowledge to these other groups.

Of the supervisory variables investigated, five appeared to be related to empathetic ability to such extent as to be considered of practical importance in the prediction of this ability. The intercorrelations of these five variables are shown in Table 3. The multiple correlation between the variables shown and the supervisors' predictions of management's responses was indicated by  $R_0 .12345 = + .53$ . The relationship between the same variables and the supervisors' predictions of labor's responses was shown by  $R_0 .12345 = + .67$ . When these correlations are compared with the correlations for a single variable of the supervisors' own scores, it is evident that this particular variable contributes the most toward the total relationship.

A Pearson  $r$  of  $+ .10$  was found between the rank order of the supervisors as rated by their superiors and their predictions for labor. The correlation between this rank-order list and predictions for management was found to be  $+ .09$ . Both of these correlations were too small to be of significance to this study. Consequently, it appears that the supervisors considered by management to be their best are able to empathize only about as well as those considered by management to be the poorest. Because of the small number of cases and the inherent weaknesses in the rank order, the results must be interpreted with caution.

Table 3

Intercorrelations of Five Variables Found to be Related to the Supervisors' Ability to Empathize

	Adaptability Scores	Age	Supervisory Experience	Education
1. <i>How Supervise?</i> Scores	$+ .48$	$-.29$	$-.23$	$+ .37$
2. Adaptability Scores		$-.44$	$-.25$	$+ .68$
3. Age			$+ .72$	$-.45$
4. Supervisory Experience				$-.32$



### Conclusions

In short, the findings indicate that the textile supervisors are unable to empathize optimally with either labor or management and that empathetic ability is related to certain psychological variables of supervision in the textile industry. The empathetic ability of a supervisor as here operationally defined can be predicted from a regression equation. The principal reason for the failure of the supervisors to understand management and labor is the projection of their own feelings, attitudes and knowledge upon these groups.

Received October 7, 1953.

### References

1. Davidoff, M. D. *A study of empathy and correlates of prejudice toward a minority group*. Unpublished Ph.D. thesis, Purdue University, Lafayette, Indiana, 1948.
2. Dymond, R. F. A preliminary investigation of the relation of insight and empathy. *J. consult. Psychol.*, 1948, 12, 228-233.
3. Dymond, R. F. A scale for the measurement of empathetic ability. *J. consult. Psychol.*, 1949, 13, 127-133.
4. Dymond, R. F. Personality and empathy. *J. consult. Psychol.*, 1950, 14, 343-350.
5. File, Q. W. The measurement of supervisory quality in industry. *J. appl. Psychol.*, 1945, 29, 323-327.
6. Libo L. M. *Attitude prediction in labor relations*. Studies in Industrial Relations No. 10. Stanford: Stanford University Press, 1948.
7. Miller, F. G. *Studies in Industrial Empathy: II. The measurement of the gap between industrial management and organized labor*. Unpublished Master's thesis, Purdue University, Lafayette, Ind., 1949.
8. Remmers, H. H. A quantitative index of social psychological empathy. *Amer. J. Orthopsychiat.*, 1950, 20, 161-165.
9. Remmers, L. J., and Remmers, H. H. Studies in industrial empathy: I. Labor leaders' attitudes toward industrial supervision and their estimate of management's attitudes. *Personnel Psychol.*, 1949, 12, 427-436.
10. Richards, A. W. *A study of some industrial relations variables in one industrial company*. Unpublished Master's thesis, Purdue University, 1950.
11. Tiffin, J., and Lawshe, C. H. The Adaptability Test: A fifteen minute mental alertness test for use in personnel allocation. *J. appl. Psychol.*, 1943, 27, 152-163.
12. Travers, R. M. W. A study in judging the opinion of groups. *Arch. Psychol.*, 1941, 260, 348-357.

## Organization Control in Business

L. R. Gaiennie

*Personnel Department, Fairbanks, Morse & Co., Chicago, Illinois*

It is becoming increasingly apparent that traditional organization charts, with their job titles and lines of authority, represent only one aspect of a business organization. They are two-dimensional still lives of a living institution and can be likened to anatomy as contrasted with physiology. Anatomy studies parts and organs of the body at rest, whereas physiology attempts to understand them in action. Organization charts fail to show actual relations between the jobs and people of an institution. This is because the various positions of a company are populated by human beings who are constantly acting and interacting to each other and to the changing conditions of business. Of course, such charts have their proper place in personnel control. Scientific personnel work was founded upon analysis of the jobs and functions of an organization. It has become commonplace to point out that job descriptions are to the personnel executive what blueprints and material specifications are to the engineer. Too often, however, the personnel executive *accepts* his organization merely because it has been formalized by charts and descriptions, and proceeds to select and train employees to fill the positions thus created.

Undue emphasis upon job relationships leads to an engineering-minded personnel administration. Such departments tend to treat people as a means rather than as an end in themselves. People are categorized as units of energy, not as people. The attempt to use people as a means rather than an end alienates them from a sense of belonging with management to the economy as we know it. Work likewise becomes a means—something foreign to a person's real interests and goals; something with which to obtain an automobile or television set; something to be given sparingly as a cost rather than a good in itself. One of the reasons why this is so lies in the fact that management, under the influence of an atomistic engineering science, has broken down its job organization in such

a way as to deprive employees of much of their creative relationship to work. The selection-minded approach to filling jobs so created has led to discouragingly meager results over the last twenty years.

It has become obvious that the personnel administrator's functions must go beyond mere analysis and acceptance of his organization. After all is said and done, personnel efficiency is measured by the success of the company and the people in it. Today, personnel administrators and psychologists are thus enlarging their field of interest—upward from the skilled hourly workers and outward toward the relationship between the positions in a given organization. The larger concept of "organization control" is enriching the older field of "employee selection." This growth and interest is indeed heartening and represents the growing maturity of both personnel and industrial psychologists. A means of relating these two aspects of industrial organization to each other is now needed if adequate organization control is to be achieved. If these two structures can be measured in similar terms, then some progress may be made, since progress in any field is largely dependent upon quantifying the data under consideration.

To perform his function effectively, the personnel executive must take a critical look at both the "job structure" and the "people structure" of his company. Both functions and people within the company must combine to produce harmony and profits. Thus, business organization has at least two structures: (1) the make-up and relationship between the various positions of the company; and (2) the make-up and relationship between the various persons occupying these positions. As has been pointed out, these two structures are merely separate aspects of the same problem. Complete understanding of each is dependent upon understanding the other, and the purpose of personnel control is to achieve proper job and personnel struc-

tures and assure a balance between them. Organization control is an attempt to measure degrees of conformity between the job organizational requirements, and the abilities and performance of job incumbents. Organizational efficiency is, in this respect, equivalent to the cost accountant's measure of standard versus actual performance. In this sense, organization control is designed to lead toward remedial or preventive action and as such lifts the older concept of employment personnel to a new and more dynamic level.

For some time, various rating techniques have been used for measuring the relative complexity of jobs as a means of relating all jobs to one another. These techniques, when applied to management positions, are usually termed "position evaluation" and form the basis of most modern salary administration. Considerable work has been done to establish the reliability of such data and, in general, it has been accepted by employees and management as the most rational and objective basis so far developed for measuring the relative value of jobs. In order to achieve our objective of relating the two aspects of organization one to the other, evaluation elements which can be applied with equal ease to both jobs and people must be used. Examples of such elements are: planning ability, skill with people, job knowledge, quality of work, responsibility, and experience. All the elements (factors) are defined in a manner equally applicable to both the job and the job incumbent, as are the various grades within each element.

By applying the usual rules of job evaluation to all of the management positions, data are obtained. After this has been done, each job incumbent is rated for the same elements against the ratings for the job he occupies. Evaluation of job incumbents is performed in a completely separate series of rating sessions. The same or different people may be used. The same principles are applied in evaluating people as are used in evaluating the jobs.

The evaluating sessions for job incumbents differ from the position evaluation series only in that: (a) different evaluators may be used for the two sessions; (b) one series evaluates people and the other evaluates positions; and

(c) the incumbent is rated against the requirements for the position he occupies.

In the studies so far made, the data from the two evaluations have been entered on master cards. These measures are then treated statistically to arrive at total scores for each job and each job incumbent. Comparisons between positions and people can then be made and it can readily be seen if an employee exceeds, equals, or is beneath the job requirements. These techniques are subject to all of the limitations and errors of any rating procedure.

When similar cards have been filled out for all employees and positions, an almost infinite number of comparisons can be made. Some of the most obvious are:

1. *Comparisons between jobs.* This can be used for such purposes as to obtain a better organization, increase or decrease the job content of certain positions, or to establish a salary structure.

2. *Comparisons between people.* Since all the people have been evaluated on the same basis, direct comparisons can be made.

3. *Comparisons of jobs and people.* This information can be used for training, organization control, upgrading, and standardization of psychometric tests.

This approach highlights the fact that reduced variances between the requirements of the job and the abilities of job incumbents can be achieved in several ways; namely:

1. By modifying job content. Where a discrepancy exists between the job and the incumbent it is possible to add or subtract duties and responsibilities.

2. By changing personnel. Balance between job and personnel can be brought about through training or transfer of personnel to achieve maximum use of company manpower.

3. By changing both job content and personnel.

If the evaluation data for positions and people thus obtained are charted, the resulting series of curves demonstrate in quantitative terms the organization as a whole. That is to say, it now becomes possible to study both the groups of jobs and the groups of people, since all have been reduced to common denominators. Curves of this kind are de-



veloped by arranging the positions of the organization in their order of complexity. It is then possible to plot both personnel and position ratings, keeping the data arranged in the order of the total position evaluations.

Separate charts for each evaluation element are also possible. For purposes of the present discussion, the following terms are defined:

1. *Position Gradient*—that curve obtained by listing the positions in ascending order of their total position evaluation scores along the abscissa and plotting their total or element position scores on the ordinate.

2. *Personnel Gradient*—that curve obtained by listing the job incumbents in ascending order of their total position evaluation scores along the abscissa and plotting their total or element personnel evaluation scores on the ordinate.

3. *Positive Variance*—any portion along a total or element evaluation curve where the abilities of the job incumbent are judged to exceed their job requirements.

4. *Negative Variance*—any portion along a total or element evaluation curve where the job requirements exceed the abilities of the job incumbent.

### Hypotheses for Experimental Test

Besides practical value to personnel executives, this method of analysis, in spite of its limitations, has certain advantages to those interested in organizational theory because of its quantitative nature. An almost infinite number of relationships between the personnel and position data can be isolated and made the subject of more detailed study. Because there has been so little research in this field, the following suggestive questions and hypotheses are stated as a means of stimulating further work on these and related organizational problems:

1. *Can selective devices, such as standardized tests, be developed using job requirements as the criterion?* The typical test standardization process in industry has been to select a group of employees working on related jobs; to separate the good from the poor performers and then to standardize the tests on this criterion. Two major weaknesses are inherent in this approach: (1) the groups used

in such a process tend to be small, thus reducing the reliability of the data; and (2) ability measures do not necessarily measure employee performance due to such factors as motivation. By using job evaluation elements such as "planning ability" as the criterion, whole populations can be tested and used in standardizing tests. This approach allows for cutting scores by job type and separates out the motivational aspect from the testing devices for separate measures.

2. *Is it better to place individuals in positions which just equal, exceed, or are less than their abilities?* This question strikes more directly at the problem of motivation and related problems. Using data obtained through this technique, it is possible to segregate separate populations from one or more organizations as follows: (a) greatly exceed job requirements; (b) exceed job requirements; (c) equal job requirements; (d) beneath job requirements; and (e) greatly beneath job requirements. Having isolated the groups for study, various experimental designs can be used to ascertain their relative efficiency. If desired, it is also possible to segregate still other subpopulations within each one of the above groups. For example, those who exceed job requirements could be subdivided into those who equal, are beneath, or are above the job requirements on a particular element.

3. *What are the effects of personnel reversals upon organization performance and morale?* There is evidence that some of the most uncooperative union stewards are antagonistic because they are more capable than the foreman over them. Applied to management organization, what are the practical results of such a situation? Is it the same at all levels of an organization? The question of reversals and their effect upon organization efficiency should probably be studied at three points on the curves: (1) those who represent reversals; (2) their superiors; and (3) their subordinates.

4. *Is it possible to have an efficient organization without positive position and personnel gradients?* In recent years, there has been considerable discussion regarding democracy in business, expressed by such terms as "bottom-up management." Is this feasible as

applied to the two organizational structures herein discussed? The proponents of bottom-up management, to the extent that they would modify personnel or position gradients, have an opportunity to study various type gradients and to report the relative efficiencies of each.

5. *Given a particular set of conditions, such as size of organization, type of activity, etc., are there particular gradients which return optimal results?* If positive answers can be given to this question, businesses might be spared thousands of dollars in cost as they establish or reorganize their operations. There is some evidence that certain generalizations may be discoverable. For example, the writer has heard competent business executives express the following ideas, which are open to experimental test under the method outlined in this paper:

- a. Job-shop organizations require more complex position and personnel gradients than production organizations.
- b. Large organizations tend to develop organization gradients significantly different from small organizations.
- c. "Mental" organizations or departments such as engineering, research, and development demonstrate flat gradients as compared to the typical manufacturing line organization.

6. *Are there optimal gradients which should be established as objectives if it is anticipated that a given organization is going to expand or contract?* The organizational strains due to change are especially apparent during quick expansion or after continued long-term growth of a company. When this happens, previous methods and personnel must adjust to the new situation. The following hypotheses are suggested as being subject to experimental verification:

- a. In a given business organization, if the job gradient expands and moves quickly upward on the ordinate, severe organizational strains will occur unless the personnel gradient is caused to do likewise.
- b. Organizational strains will ensue if the job gradient changes its relative shape while the personnel gradient remains the

same. Empirically speaking, it is probable that production control installations have a high mortality rate because they are frequently installed by outsiders who convince top management their system can be installed without disturbing existing personnel. When this happens, the new system frequently creates severe organizational strain (modifying the job but not the personnel gradient; thus establishing negative variance) or else existing personnel through mass inertia finally defeat the new system.

7. *Given a particular organization, are there optimal organizational curves which relate to particular company policies and procedures?* Observations to date indicate that such may be the case. For example, highly centralized multi-plant companies display different organizational gradients than highly decentralized multiple plant operations. Certain corporations which have centralized or decentralized their organizations in recent years are known to have created organizational strains which might have been reduced if they had considered their existing gradients in relation to proposed objectives before starting their programs.

8. *Can training programs be made more realistic and be given to those employees who need assistance in the particular problem areas uncovered?* Work so far indicates that much of the training time spent in industry is of a blunderbuss variety. Companies are too often prone to dangle a watch in front of bored employees in the hope that such sessions will somehow improve performance. This technique allows for selective training or transfer of employees based upon measures taken. In addition, it allows for post-training measures to ascertain the relative effectiveness of such programs.

The above questions and hypotheses are meant to be suggestive of the kinds of problems which can be attacked experimentally through use of the position and personnel evaluation method. Further work on these and related problems is badly needed to establish factual guides for the business executive.

*Received October 7, 1953.*

# Quantitative Analysis of Verbal Evaluations \*

Sidney H. Newman

*U. S. Public Health Service, Washington 25, D. C.*

As ordinarily used, verbal evaluations included in performance reports furnish "impressions" and "qualitative observations." Quantitative analysis of such comments can increase the usefulness of these reports and establish a reliable basis for comparing individuals. This paper describes the development and reliability of a procedure for scoring comments obtained from an efficiency report used to evaluate the job performance of commissioned officers in the United States Public Health Service. This work is part of the research program discussed by Newman (4).

## Method

The quantitative method developed here for the analysis of verbal evaluations involved the adaptation of three well-known techniques.

First, the method of content analysis suggested the classification of supervisors' comments into categories. Content analysis has been employed extensively by Lasswell and his associates (3) in analyzing the political and propagandistic content of mass media.

Second, the technique evolved by Thurstone (7) for scaling attitudinal statements was used to assign values to comments classified in each of the categories. Other investigators such as Uhrbrock (8) have applied the Thurstone technique to the scaling of statements concerning job performance and personal characteristics.

Third, methods like those introduced by Thorndike (6) for measuring the quality of handwriting were the basis for the use of a master scale for scoring each comment.

The procedure used in establishing a system for scoring the comments in the efficiency report consisted of the following steps:

1. A total of 779 comments were collected from the "remarks," "handicaps," and "rec-

ommendations" sections of several hundred officer efficiency reports. A comment was defined as any word, phrase, or clause constituting a unitary evaluative description of the officer upon whom the report was prepared.

2. By sorting and grouping all comments, it was possible to establish 12 descriptive categories relevant to officer characteristics deemed important to the Service and sufficiently independent of each other to allow

Table 1

Reliability Coefficients for Scale Placements  
in Each Category

Category	No. Items	$r_{11}$ *	$r_{111}$ †
General evaluation	69	.90	.99
Potentiality for future	32	.86	.99
Training and experience	30	.87	.99
Relations with work associates	35	.85	.98
Relations with official groups	16	.93	.99
Relations with patients and public	26	.95	.99
Motivation	39	.88	.99
Job proficiency	70	.89	.99
Job progress	23	.89	.99
Potentiality as candidate for Regular Corps	23	.89	.99
Work attitudes	60	.86	.99
Intellectual qualities	88	.95	.99
Intellectual qualities ("duplicate" items removed)‡	46	.93	.99

\*  $r_{11}$  = Average of correlation of each judge's placements with every other judge's placements.

†  $r_{111}$  = Reliability of average of 11 judges.

‡ To test the hypothesis that the high correlations were produced by "duplicate" items, these were removed. "Duplicate" items were defined as those having: (a) the same adverb modifiers; and (b) scale placements differing by no more than one place. It may be seen that removing these items has very little effect on the correlation.

\* The writer wishes to acknowledge with appreciation the aid of Mrs. Jane S. Harris who was Scorer 1 and who also did much of the statistical computation.



classification of comments. These categories are listed in Table 1.

3. The comments in each of the categories were placed on a nine-point scale from "undesirable" (1-3), to "neutral" (4-6), to "desirable" (7-9) by 11 Public Health Service judges, most of whom were psychologists. Each comment was assigned a numerical value for use in scoring; this value was the median of the scale placements made by the different judges.

4. A scoring manual was constructed by listing, in each of the 12 categories, the comments with their median values. In using the manual, the scorer identifies a comment and classifies it in one of the 12 categories. This process of identification and classification is defined here as coding. He then matches each comment as closely as possible with one in its category in the manual and assigns it the listed numerical value. The numerical values of all comments in an efficiency report are averaged to obtain the raw score for the verbal evaluation parts of the report. In this article, the entire process of arriving at scores, involving both the coding of comments and assigning of numerical values to them, is termed scoring.

5. Reliabilities of the scale placements and the scoring methods were determined.

Results

*Reliability of Scale Placements.* As obtained by the method of average intercorre-

lation, Peters and Van Voorhis (5), the reliabilities of the scale placements made in each of the 12 categories by the 11 judges are shown in Table 1.

For comparative purposes, split-half coefficients based on correlations of the averages of five judges with those of five other judges and stepped up for 10 judges by the Spearman-Brown formula were computed for four categories. The results for 10 judges were similar to those obtained by the method of average intercorrelation for 11 judges (see Table 1): training and experience,  $r_{55} = .95$ ,  $r_{1010} = .97$ ; relations with work associates,  $r_{55} = .98$ ,  $r_{1010} = .99$ ; relations with patients and public,  $r_{55} = .99$ ,  $r_{1010} = .99$ ; proficiency,  $r_{55} = .98$ ,  $r_{1010} = .99$ .

*Reliability of the Scoring Method.* The reliabilities of the comment scores assigned under various conditions are presented in Table 2.

In one method of determining reliability, two people, one trained by the other, independently scored the comments in a sample of officer efficiency reports. The scores assigned by the different scorers were correlated (Scorer 1 vs. Scorer 2).

In the other method, scores assigned by the same scorer on two different occasions were correlated (Scorer 1 vs. Scorer 1, and Scorer 2 vs. Scorer 2). Precautions were taken to minimize the effects of memory and other contaminating factors. During the period intervening between the two scorings, the scorer

Table 2  
Reliability of Comment Scoring

	No. Efficiency Reports Scored	Total No. of Different Comments Coded in Both Scorings	Comments Coded the Same in Both Scorings		Correlation Between Scores on Comments Coded the Same in Both Scorings $r_{11}$
			No.	Per Cent	
Scorer 1* vs. Scorer 2	32	114			
Scorer 1 vs. Scorer 1 (4 mo. interval)	30	99	99	87	.86
Scorer 2 vs. Scorer 2 (4 mo. interval)	39	141	93	94	.95
Scorer 2 vs. Scorer 2 (14 mo. interval)	40	151	115	82	.95
			113	74	.96

\* Scorer 1 was more experienced in scoring than Scorer 2.

worked on other efficiency reports and was not allowed to see those utilized in the studies. The extent of agreement in coding comments in both of the scorings is also shown in Table 2 for all four studies.

In the comparison of the scoring done by Scorer 2 on two occasions, separated by a 14 month interval, the comment scores assigned were also averaged to obtain a single score for the verbal evaluation parts in each efficiency report. The average scores assigned each of the 40 reports on the two different occasions correlated .94.

### Discussion

A procedure for quantitatively analyzing verbal material for the comparison of officer performance has been developed. The placements of verbal comments on a nine-point scale, basic to the development of the scoring procedure, can be reliably achieved by a relatively small number of judges. In agreement with these findings, Uhrbrock (8) obtained high reliabilities for the Thurstone scale values of descriptive rating scale statements; Hinckley (2) and Ferguson (1) found that scale values assigned attitudinal statements by use of the Thurstone technique were highly reliable.

Three aspects of the reliability of scoring by the use of the scoring manual were considered: the correlation between scale values assigned the comments, the correlation between average comment scores, and the percent of agreement in the coding of comments.

The reliability of comment scores was found to be higher for scores assigned by the same scorer on two different occasions (.95 in one study and .96 in the other) than for the scores assigned by two different scorers (.86). Increasing the lower coefficient might be accomplished by adding the occasional new comments to the large sample in the manual, and by increasing the similarity of judgmental standards through cooperative training and scoring.

The reliability of average comment scores is analogous to the reliability of total scores on a test. In one of the studies (Scorer 2 vs. Scorer 2, separated by a 14 month interval),

it was found that the reliability of average comment scores (.94) was similar to the reliability of individual comment scores (.96). This suggests that when the reliability of comment scores is found to be high, average comment scores will also be reliable.

Agreement in the coding of comments may be considered fairly high, since in three out of the four studies the percentages of agreement were, respectively, 82, 87, and 94 per cent. In the remaining study, the percentage of agreement was 74 per cent, but this coding was done by the more inexperienced scorer, with a 14 month interval between the scorings. This interval probably represented a training period in which the scorer may have developed more ability to code comments. Agreement in the coding of comments might be increased, in general, by giving scorers extensive training in this aspect of the method.

The findings on reliability of the procedures used here for scoring comments in efficiency reports suggest that this method may be useful for analyzing quantitatively other kinds of verbal material. In occupational situations, supervisors usually like to make verbal reports; this scoring procedure will allow quantitative utilization of material which has ordinarily been merely "taken into consideration." It is also likely that these quantitative procedures will prove useful in such fields as propaganda analysis, the analysis of literature, or the analysis of the verbal reports of patients or interviewees. Of course, it would be necessary to determine the reliability, validity, and other relevant characteristics of the scores obtained in any given situation.

### Summary and Conclusions

A method for the quantitative analysis of verbal material is presented. Verbal material is quantified by categorizing each unit of material (each comment) and comparing it with the empirically derived master scoring scale constructed for that category. The findings show that: (a) the comments in each category can be reliably placed on a nine-point scale by a relatively small number of judges; and (b) scores, based on either the individual comments in each category or the average

comment scores for each report, are reliable. It is suggested that the procedures developed here can be utilized for other types of verbal material.

Received October 26, 1953.

### References

1. Ferguson, L. W. The influence of individual attitudes on construction of an attitude scale. *J. soc. Psychol.*, 1935, 6, 115-117.
2. Hinckley, E. D. The influence of individual opinion on construction of an attitude scale. *J. soc. Psychol.*, 1932, 3, 283-296.
3. Lasswell, H. D., Leites, N., and associates. *Language of politics; studies in quantitative semantics*. New York: G. W. Stewart, 1949.
4. Newman, S. H. The officer selection and evaluation program of the U. S. Public Health Service. *Am. J. Publ. Hlth.*, 1951, 41, 1395-1402.
5. Peters, O. D. and Van Voorhis, W. R. *Statistical procedures and their mathematical bases*. New York: McGraw-Hill, 1940.
6. Thorndike, E. L. Handwriting. *Teacher's Coll. Rec.*, 1910, 11, 83-175.
7. Thurstone, L. L. Attitudes can be measured. *Amer. J. Sociol.*, 1928, 33, 529-554.
8. Uhrbrock, R. S. Standardization of 724 rating scale statements. *Personnel Psychol.*, 1950, 3, 285-316.



## Standardization of the GATB for the Occupation of Tabulating Machine Operator<sup>1</sup>

Minnesota State Employment Service in Cooperation with the U. S.  
Employment Service, U. S. Department of Labor,  
Washington, D. C.

This study is concerned with the prediction of success or failure in the occupation of Tabulating Machine Operator. It was conducted by the Minnesota State Employment Service in cooperation with the United States Employment Service (USES) and the National Machine Accountants Association (NMAA).

The present study is an attempt to develop national norms for this occupation. It is an outgrowth of previous studies conducted in Florida, Ohio, and Minnesota, the latter being conducted in cooperation with the Northwest Chapter of NMAA and the University of Minnesota.

### Procedure

*The Sample.* The sample in this study is composed of 203 operators employed in four states, viz., California, North Carolina, New Jersey, and Wisconsin. Of the 203 operators, 96 are women and 107 are men.

All types of operations listed by the International Business Machines Company (IBM) and by Remington Rand were performed by operators in the sample. Participating firms were instructed to refer all tabulating machine operators for testing. If this procedure was not feasible,

operators were to be selected for testing who were representative of operators employed by the firm with respect to age, sex, work-level, and experience.

All operators had been employed for six months or longer so that they had completed the probationary period for this occupation.

*The Criterion.* The criterion was a rating scale<sup>2</sup> which included items considered by selected Tabulating Machine Supervisors to be important for successful work performance as a Tabulating Machine Operator.

Supervisors were instructed to rate operators in comparison with Tabulating Machine Operators "in-general." This instruction was used to obtain, as nearly as possible, comparability of ratings among the participating firms. A re-rating was conducted within a two-week period for the purpose of determining reliability. A reliability coefficient of .878 with a standard error of .004 was obtained. Since re-ratings were not available for the entire sample, the first rating was used as the criterion.

The rating scale was composed of 8 items for which the rater had five choices of response indicating the degree of performance of the operator. Weights of 1 through 5 were assigned to these responses so that the minimum possible score was 8 and the maximum was 40. The mean score was 26.05 with a standard deviation of 6.7 and the range was 8 through 40 for the sample of 203 operators.

All operators having scores one standard deviation below the mean, or lower, were placed in the Low criterion group. Therefore, 37 operators comprise the Low group and 165 operators were contained in the High criterion group.

*The Predictive Instrument.* The machine-scorable form of the General Aptitude Test Battery (GATB) was used for this experimental study. This battery, composed of 12 tests, measures 9 aptitudes, viz., general intelligence (G), verbal ability (V), numerical ability (N), spatial aptitude (S), form perception (P), clerical aptitude (Q), motor coordination (K), finger dexterity (F), and manual dexterity (M). The general

<sup>1</sup> Ruth E. Potter, State Test Technician of the Minnesota State Employment Service, had major responsibility for the supervision of the total study and the preparation of this article. Participating in the development of the experimental design for the study were Ruth E. Potter and John R. Boulger of the Minnesota State Employment Service; Dr. Beatrice J. Dvorak and Albert Mapou of the United States Employment Service; and Mr. Wayne Spielman of the National Machine Accountants Association. At the Minnesota State Employment Service, James Ryan and Robert Coll conducted the statistical analysis of the data. At the national office of the United States Employment Service in Washington, D. C., the following persons participated in the planning of the study, coordination of the collection of data by the New Jersey, North Carolina, Wisconsin and California State Employment Services, and review of the completed study: Dr. Beatrice J. Dvorak, Albert Mapou, Charles Meigh and Sylvia Hoke. For the National Machine Accountants Association, Mr. Wayne Spielman directed the promotional activities among NMAA membership.

<sup>2</sup> The Staff also wishes to acknowledge the impetus given the study by Kenneth Schenkel whose Ph.D. research was the 1952 study. Through Dr. Schenkel came the contacts with the NMAA and, apart from the standard USES materials and approach, the rating scale and accessory materials (modified) developed by him were used for this study.

working population norms are established on the basis of a selected sample of 4000, stratified to obtain proportional occupational representation as shown by the 1940 Census of the Population. The general-population means for aptitudes in the battery are 100, with standard deviations of 20.

*Statistical Analysis.* The significance of GATB aptitudes for the occupation of Tabulating Machine Operators was determined on the basis of mean aptitude scores, standard deviations, validity coefficients, and job analysis data, as shown in Table 1.

Results

Aptitudes significantly related to success in the occupation as evidenced by high mean scores, low standard deviations, significant validity coefficients and identification through job analysis are: (G) general intelligence; (N) numerical ability; and (Q) clerical aptitude. Spatial aptitude (S) is also related to the occupation as indicated by the validity coefficient, identification through job analysis, and because it adds to the selective efficiency of Aptitudes G, N, and Q.

Minimum scores for Aptitudes G, N, Q, and S were set approximately one sample standard deviation below the sample mean rounded to the nearest five-point score level. This results in norms consisting of G-95, N-95, S-85, and Q-100.

To evaluate the selective efficiency of these norms in terms of the relationship between

Table 1

Means (M), Standard Deviations (S.D.), Pearson Product-Moment Correlations with the Criterion (*r*) for the Aptitudes of the GATB

Note: N = 203 Tabulating Machine Operators.

Aptitude	M	S.D.	<i>r</i>
G	111.4	14.4	.34**
V	109.1	15.1	.22**
N	111.6	14.8	.36**
S	106.5	18.3	.20**
P	109.9	13.9	.10
Q	116.4	15.1	.15*
K	112.0	16.4	.08
F	105.6	19.9	.10
M	106.7	20.9	.10

\* Significant at the 5% level.  
\*\* Significant at the 1% level.

Table 2

Relationship Between Pass-Fail on Test Norms Consisting of Aptitudes G, N, S, and Q with Critical Scores of 95, 95, 85, and 100, respectively, and the Criterion

Group	Fail	Pass	Total
High	40	126	166
Low	20	17	37
Total	60	143	203

$r_{tet} = .48;$        $\sigma_{tet} = .14$   
 $\chi^2 = 11.643,$        $p/2 = .001$

those operators passing and failing the norms and those in the High and Low criterion groups, tetrachoric correlation and Chi-square techniques were employed. The relationship between test norms and the criterion is shown in Table 2.

Both the Chi-square test and the tetrachoric correlation indicate a statistically significant relationship between passing the test norms and success on the job, as measured by the criterion. Fifty-four per cent of the Low criterion group fail the norms, while 76 per cent of the High group pass the norms.

*Cross-Validation.* Previously derived norms based on the original Minnesota sample, and samples of independent studies conducted in Ohio and Florida were applied to the national sample. Although these norms were related to job success, they were not as predictive of job success for the national sample as they were for the samples from which they were derived. In general, the same aptitudes appeared to have predictive value for each of the studies, but some variation was found with respect to the critical scores obtained.

Summary

This study reports the development of national norms, based on the GATB, for the occupation of Tabulating Machine Operator.

General intelligence, numerical aptitude, spatial aptitude, and clerical ability were found to be significantly related to success in the occupation.

Received September 10, 1953.



## Comparative Validities in Clerical Testing

Edward N. Hay

*Edward N. Hay & Associates, Inc., Philadelphia, Pa.*

In connection with another project the Hay Number Perception Test and the Wonderlic Personnel Test were administered to 19 candidates for a special task in a life insurance company. It was observed that there was an extraordinary number of high Personnel Test scores. Out of 19 people tested 7 had scores of 40 or more, which is above the 98th percentile in similar groups. The mean for this group was 35.6; about the 95th percentile. This insurance company had been using the LOMA No. 2A test, which is designed for the selection of clerical workers. A correlation of .66 was found between Personnel Test and LOMA No. 2A, but an  $r$  of only  $-.01$  between Personnel Test and Number Perception Test. Thus LOMA No. 2 test seemed to be excluding any except fairly bright applicants.

However, it is known that the LOMA No. 2A test is a good predictor of success in simple routine clerical work, as well as of promotability. The Number Perception Test has also established its efficiency in routine clerical selection and consistently correlates very low with mental ability tests. The question immediately presented itself as to which of the two tests, LOMA No. 2A or Hay Number Perception, would show the higher validity in this company in predicting speed of production for *low-level* clerks. This first group afforded no criterion of success, since it was a mixed group with the majority of the employees in supervisory or technical positions. So, another group, engaged in simple routine clerical work, was selected in order to compare the validity of these two tests. The SRA Clerical test was also available so it was administered, too. The subjects were the 24 clerks in one department, all but two in the lowest pay classifications and performing simple routine tasks. Of these, 23 were women and none had had any supervisory responsibility. Average length of service was 37.1 months, with six over 5 years, nine be-

tween 1 and 5 years, eight under 1 year and one at five months. Correlation between length of service and the supervisor's ratings described below yielded a coefficient of  $-.08$ .

*The Tests.* LOMA No. 2A is a test available only to life insurance companies. It is an omnibus work-limit test in six parts: checking, directions, same-opposites, proverbs, arithmetic and spelling. Score is a combination of time and errors. Administration time averages about 35 minutes.

*Wonderlic Personnel* test is a well-known mental ability test composed of a variety of verbal and numerical problems.

*SRA Clerical* is in three parts, speeded and timed separately. Vocabulary is a 5-minute test of 48 items. Arithmetic allows 15 minutes for 24 problems of numerical reasoning. Checking is a 5-minute coding test of 144 items.

*Hay Clerical Battery* is composed of three speeded tests of 4 minutes each. Number Perception has 200 pairs of three- to six-digit numbers, the task being to check those that are the same. Name Finding requires the subject to look at a name and remember it well enough to pick it out of a group of four similar names on the back of the sheet. Number Series consists of 30 simple number series completion problems.

*The Criterion.* The criterion was the average of the ratings made by the department head and assistant department head. They were made about three weeks apart and wholly independently. The rating method employed three rating principles in combination; graphic scale, man-to-man comparison and forced distribution. All 24 names were listed on a single rating sheet described as "Speed of Working" and the rater was asked to place a check mark on the line opposite each employee's name in such a way that approximately one-half of the names were checked in a vertical band designated as "Average," about one-fourth "Above De-



partment Average" and about one-fourth "Below Department Average." Distinctions among employees in each of these three groups were to be indicated by the relative positions of the check marks along the lines. After the ratings were completed the value of each mark was measured on a scale ranging from 0 to 40, this particular scale being arbitrary.

The product-moment correlation between the two sets of ratings yields an  $r$  of .89, indicating a highly reliable criterion.

### Results

The second column in Table 1 shows the correlations between scores on the various tests and the average of the scaled values of the two ratings. These coefficients point to the greater efficiency of four of the tests, but such coefficients cannot always be relied upon, especially in so small a sample, because regression is not always rectilinear.

The third column of Table 1 shows the best cutting score on each single test or combination of tests. The 24 cases fell into five groups as rated by the two supervisors.

Group I was rated "Good" by both raters; group II was rated "Good" by one rater and "Average" by the other, etc. The first three groups were considered "Good." Groups IV and V were considered "Poor" since one or the other rater had so classified all members. Cutting scores were selected by inspection which would admit the greatest number of subjects rated "Good" to that group and exclude the greatest possible number rated "Poor."

The only combination of tests which would increase predictive efficiency was Number Perception and Name Finding, which correctly assigned 21 out of 24 subjects to the proper group, "Good" or "Poor." This was significantly better than chance at the one per cent level.

### Discussion

This study confirms other similar studies with some of the same tests in showing that prediction of success in low-level routine clerical work is usually more efficiently accomplished by tests based on what appears to be speed of perception than by tests in-

Table 1  
Predicting "Speed of Work" from Test Scores  
N = 24

Test	Time, Min.	<i>r</i>	Best Cutting Score	Correct Selection <sup>1</sup>	Signifi- cance
Wonderlic Personnel	12	.04	22	16 of 24	No
SRA Clerical: Vocabulary	5	.08	29	15 of 23	No
Arithmetic	15	-.05	13	15 of 23	No
Coding	5	.55	72	18 of 23	.05
Hay: Number Perception	4	.64	115	18 of 24	.10
Name Finding	4	.60	19	18 of 24	.10
Number Series	4	.04	19	18 of 24	.10
LOMA No. 2A <sup>2</sup>	35	.54	89	18 of 24	.10
Hay: Number Perception } + Name Finding }	8	.67 <sup>3</sup>	115	18 of 24	.01
SRA Clerical: Vocab. }	25	.55 <sup>3</sup>	19	21 of 24	.05
Arith. }			29		
Coding }			13	18 of 23	
			72		

<sup>1</sup> Chance would give a correct selection of 12 out of 24. Perfect selection would be 24 out of 24.

<sup>2</sup> No correction has been made for possible restriction of range due to the use of this test in original selection. Range in sample however was as great as for other tests given, judging by reference to published tables of norms.

<sup>3</sup> Multiple  $R$ .

volving primarily reasoning problems (1, 2, 3, 4, 5, 6, 7, 8).

It is worthy of note that the most efficient tests were also the briefest: Number Perception, Name Finding and SRA Checking. This points to the wasted effort of giving a large battery of tests, tests with long time requirements or an omnibus test, where some material may be only dead wood and may even reduce the efficiency of the whole test. Time is not very important in school situations but in industry it is critical, both for maintaining good public relations and in reducing the direct costs of testing.

Warning has already been given against placing complete reliance on product-moment coefficients of correlation, on the ground that if regression is not rectilinear the coefficient may thereby be lower than would be expected. Table 1 affords an example. Number Series correctly selects 18 out of 24 cases, the same figure achieved by three other tests and nearly as high as a fourth; yet the  $r$  is only .04, whereas the others are between

.54 and .64. An examination of the scatter-diagrams provides the explanation: regression follows a U-shaped course for Number Series but is almost perfectly rectilinear for the other four tests.

Received May 28, 1954.

Early publication.

### References

1. American Bankers Association, New York. *Clerical Testing in Banks*. 1952.
2. Blakemore, Arline. Reducing typing costs with aptitude tests. *Personnel J.*, 1951, 30, 20-24.
3. Hay, E. N. Predicting success in machine book-keeping. *J. appl. Psychol.*, 1943, 27, 483-493.
4. Hay, E. N. Cross-validation of clerical aptitude tests. *J. appl. Psychol.*, 1950, 34, 153-158.
5. Hay, E. N. Test scores and ratings of clerks at the Roane-Anderson Co. Unpublished, 1950.
6. Hay, E. N. Mental ability tests in clerical selection. *J. appl. Psychol.*, 1951, 35, 250.
7. Howe, D. W. Summary of test validation studies at the Hanover Bank. Unpublished, 1950.
8. Miller, R. B. Reducing the time required for testing clerical applicants. *Personnel J.*, 1950, 28, 364-366.

## A Sales Comprehension Test \*

Martin M. Bruce

*Dunlap and Associates, Inc., Stamford, Conn.*

One of the areas in industry where testing has been carried on extensively is sales. However, very few instruments dealing with selling have been published and are available for general distribution.

The reader will find an excellent review of the literature on selection of sales personnel in Husband's article (3). Since that publication Rock (4) has published an article on his Sales Situations Test. Rock reported on just two small sales groups in describing the test, one consisting of 25 subjects, the other 31 subjects. The instrument attempts to present "live" situations, with items in multiple choice format. The idea has long appeared to the writer to be a sound one.

Because of the apparent dearth of testing material in a field where a great many men are tested, the writer set about in 1946 to devise a test that would aid in measuring potentiality for success in selling.

### Problem

The problem was one of constructing a test that would aid in predicting success in selling. Specifically, the test was to be one that would be directly applicable to the wholesale sales field in general.

### Procedure

In 1946 an experimental form of the test was prepared in mimeograph format. This instrument contained 74 items. The items were constructed with the aid of salesmen in various fields, business men in occupations related to selling, industrial psychologists, and literature on selling.

The 74 items were administered to salesmen in various fields throughout the country as well as to individuals in occupations other than sales. The 50 items that differentiated best between the sales and non-sales groups were retained and published as the *Aptitudes Associates Test of Sales Aptitude (Principles of Selling) Form A (2)*.

\* *The Sales Comprehension Test, Form M* by Martin M. Bruce is obtainable from the author at 71 Hanson Lane, New Rochelle, New York.

Additional data on 1,404 cases were collected on the 50-item form. These cases consisted of 1,007 non-salesmen and 397 salesmen. The non-salesmen consisted of individuals applying for all types of jobs other than sales with companies in the East and Midwest, students studying psychology in New York and New Jersey colleges, vocational guidance clients in New York City and men in various non-sales jobs throughout the country. The sales group consisted of 55 salesmen of major and small electrical appliances in cities in Ohio and Connecticut; 86 salesmen and sales managers of electronics products located in practically all common distribution centers in the United States; 19 metropolitan New York salesmen of office dictating equipment; 13 salesmen of hardware products located in Southern and Midwest locations; and 224 other individual salesmen in a wide variety of fields located in all sections of the country. This last group included salesmen in the following fields: office supplies, whiskey, beer, soap, razor blades, fountain pens, automatic pencils, clothing, textiles, furniture, dairy products, advertising space, pharmaceuticals, books, materials handling products, machinery, and a number of others. Phi coefficients were computed for the 50 items on the basis of the above samples. The 30 "best" items were retained and published as the *Sales Comprehension Test, Form M*.

A cross-validation study was conducted by administering the 30-item form to 661 additional non-salesmen and 334 salesmen. The non-salesmen were in 22 different states and filled 21 different jobs. The salesmen were employed in 18 different states and were employed in 11 different sales fields.

An additional validity study was conducted with a group of 82 sales managers employed throughout the United States by a door-to-door cosmetics sales firm. These sales managers supervise a group of full and part time saleswomen who sell on a commission basis.

### Results

**Validity.** Computations for the sales and non-sales groups containing the original 397 and 1,007 cases, respectively, yielded a  $t$  of 13.1. This finding suggests that there is less than one chance in 100 that the means of these samples are not significantly different. However, this measure of difference is spuriously high since it is based on the same



population from which the Phi coefficients were computed. The means were 30.8 for the sales population and 11.1 for the non-sales population. The SD's were, respectively, 13.8 and 18.7.

The cross-validation populations of 661 non-salesmen and 334 salesmen yielded a  $t$  of 5.8. This statistic brings us beyond the 1% level of confidence in assuming that the sales and non-sales populations are not similar in their responses on this test. This is an indication of the test's status validity (5). Overlapping amounts to 19% in these populations, this percentage of the non-sales group equalling or exceeding the median of the sales group. In this cross-validation population the means of the sales and non-sales groups were, respectively, 28.9 and 12.2; the sigmas were 12.2 and 16.9; the medians 29 and 12.

In a study conducted with the sales force of a nation-wide electronics sales firm six tests were completed by the 86 salesmen and sales managers. These included personality inventories, mental ability and other ability tests and an interest inventory. The Sales Comprehension Test correlated higher with the rating criterion than any of the other six tests employed in the battery. The  $r$  was .32. The criterion has an uncorrected odd-even reliability of .92.

In this group the mean scores for the 77 salesmen and 9 sales managers were compared by computing  $t$ . The  $t$  of 2.4 is significant at the 2% level. There is a 31% overlap here, using the same overlap measure as above. Assuming that sales managers as a group have better sales comprehension than salesmen, the indication that the Sales Comprehension Test measures this difference further suggests validity for the test.

Scores on this test were correlated with final grades of 27 students studying salesmanship at Rutgers University. The  $r$  proved to be .68, suggesting that this test measures comprehension similar to that gained by students studying salesmanship in school.

*Correlation with Intelligence.* It is a common research finding that abilities tend to be positively correlated. A particularly frequent finding is that tests employed in the same situation tend to correlate positively with

each other and especially with tests of intelligence (1). Statistical analysis usually reveals that various paper and pencil tests actually measure to a significant extent what intelligence tests measure. Therefore, it is important to know the extent to which this test is related to measures of intelligence.

A correlation was run between the total score on the Sales Comprehension Test and the total score on the Otis Self-Administering Test of Mental Ability, Higher Examination: Form A. The correlation based on a sample of 387 men, women and salesmen was  $-.19$ . This group was composed of college psychology students studying testing, job applicants and vocational guidance clients. In this group the standard deviation of Otis raw scores was 9.7 and the standard deviation of Sales Comprehension Test raw scores was 18.7. The means were, respectively, 56.7 and 19.8.

Further research was conducted with the aid of Thurstone's Primary Mental Abilities Test which contains five factors. The 173 subjects include 159 men and 14 women. All but four of the men and two of the women were evaluated for clerical, sales, managerial or engineering positions with various firms in the East.

The findings appear in Table 1.

The Sales Comprehension Test score mean and standard deviation were, respectively, 17.2 and 16.9.

The fact that all of these correlations are close to zero and since the correlation with the Otis is low and negative, it appears justified to state that measures of various intelli-

Table 1  
Relationships Between Sales Comprehension  
Test and PMA Test

PMA Factor	N	Mean	Sigma	$r$ with Sales Test
Verbal Meaning	173	40.4	7.8	.06
Space	173	24.6	9.5	-.20
Reasoning	173	17.2	4.9	-.05
Number	170	46.7	11.5	-.08
Word Fluency	170	37.0	14.2	.02
Total Score	170	224.9	47.8	-.05

gence factors and the Sales Comprehension Test are not related. The Sales Comprehension Test appears to measure something other than intelligence.

*Correlation with Persuasive Preference.* There appears to be a positive linear relationship between persuasive preference as measured by the Kuder Preference Record, Form CH and performance on the Sales Comprehension Test. Data on these two tests were obtained for 146 non-salesmen and 54 salesmen. The  $r$  proved to be .39, significant at the 1% level. The standard deviation of persuasive preference scores was 15.8 while the standard deviation of Sales Comprehension Test scores was 17.1. The respective means were 37.7 and 17.8.

The modest but positive and significant  $r$  between sales score and persuasive score is in keeping with the concept that people tend to learn in areas in which they are interested.

*Reliability.* Reliability data in the form of tests and retests were obtained from 103 college students. Scores ranged from -36 to 47. The mean of the first testing was 10.4 and for the second it was 11.1. The standard deviations were, respectively, 15.8 and 14.9. The test-retest reliability coefficient for this group was .71. Because this is a restricted group with respect to range of scores, it seems likely that the true test-retest reliability coefficient for the entire population, including salesmen, is somewhat higher. The  $r$  is .79 when corrected for homogeneity.

### Summary

An experimental form of a test to aid in selecting and evaluating salesmen was prepared in 1946. Preliminary validity data led to the elimination of 24 of the 74 multiple choice items. Over a period of five years data were collected on the 50-item form.

Data on 1,398 cases indicated that there were 30 items that significantly and reliably differentiated salesmen from non-salesmen. These items have been combined to form the Sales Comprehension Test, Form M. This test was cross validated on a supplementary population.

This instrument proved to be the most valuable in predicting success among salesmen and sales managers in a national sales organization. The test correlated significantly with final grades in a class in sales principles. The instrument, unlike other paper and pencil tests, does not measure intelligence to any extent. People who show high preference for persuasive activities tend to do better on this test. The test appears capable of differentiating good from poor sales personnel.

A test-retest reliability coefficient, .79 corrected for homogeneity, is sufficiently high for group situations to warrant confidence in its consistency of measurement.

The Sales Comprehension Test, Form M, appears to be an instrument that can be utilized in sales selection and evaluation situations. Its validated item content also lends itself to sales training situations.

*Received November 6, 1953.*

### References

1. Bruce, M. M. The prediction of effectiveness as a factory foreman. *Psychol. Monogr.*, 1953, 67, No. 12 (Whole No. 362).
2. Buros, O. K. *Fourth Mental Measurements Yearbook*. Highland Park: Gryphon Press, 1953.
3. Husband, R. W. Techniques of salesmen selection. *Educ. psychol. Measmt.*, 1949, 9, 129-148.
4. Rock, M. L. A sales situation test. *J. appl. Psychol.*, 1951, 35, 331-332.
5. Technical recommendations for psychological tests and diagnostic techniques: preliminary proposal. APA Committee on Test Standards. *Amer. Psychologist*, 1952, 7, 461-475.

## A New Method for Obtaining Weighted Composites of Ratings

H. F. Dingman and J. P. Guilford

*University of Southern California*

In spite of the many weaknesses of ratings of personnel obtained in the practical situation, they still often remain the only criterion against which to validate predictive measures. It is therefore important that we correct for weaknesses wherever we can in order to achieve the best information obtainable concerning the validity of selection instruments.

It has been amply demonstrated that in order to obtain increased reliability, and hence also probably increased validity, of criterion ratings, it pays to combine ratings from several raters. One of the common difficulties in this connection, however, is that no rater is acquainted with all the ratees in the experimental group. At best, not all raters are equally well informed concerning all ratees. It is also true, even when raters know ratees fairly well, that each rater uses different information and rates on different qualities. Under such conditions not all ratings should be given equal weight in forming composites. This report is concerned with the development of a method of weighting obtained ratings in terms of two rater characteristics. One is the rater's tendency to rate on qualities in common with other raters and the other is the rater's degree of confidence in his rating of particular individuals.

The problem of weighting ratings arose in connection with a project on the validation of a new testing instrument designed for the selection of personnel who come under the general category of Psychiatric Technicians (Ward Aides) serving in a state institution.<sup>1</sup> A total of 716 such personnel in the same institution were under study. Each one had been rated by four different supervisors who had been in positions favorable for observing their performances. A graphic rating was given on a line seven centimeters long under the instruction to rate for general effectiveness on the job. Each rater also gave a rating on a similar line indicating his own degree

of assurance that his rating of effectiveness was correct.<sup>2</sup> Some of these ratings would be zero or near zero where the raters felt that they had little or no basis for making the rating of effectiveness.

### Intercorrelations

Before adopting any system of weighting the ratings to form a composite criterion measure, we decided to obtain as much information as possible concerning the properties of the ratings. This was accomplished through intercorrelations of raters and factor analyses of both the effectiveness and the assurance ratings.

Table 1  
Intercorrelations of Effectiveness Ratings  
(N = 716)

	Rater			
	A	B	C	D
A		.54	.16	.53
B	.54		.11	.45
C	.16	.11		.08
D	.53	.45	.08	

Table 1 shows the intercorrelations among the four raters, using the effectiveness ratings of all 716 employees. It is obvious that raters A, B, and D show about the same level of inter-rater agreement on ratings of effectiveness, while rater C shows little agreement with any of those three.

The factor analysis of the correlation matrix was carried out by the centroid method, with iterative solutions until communalities were stabilized. The results appear in Table 2. Here it is seen, first, that one common factor is sufficient to account for the intercorrelations. It can also be seen that rater A has definitely the highest communality. This is significant in view of the fact that A was a supervisor who makes the major decisions concerning work assignments and in-

<sup>2</sup> The suggestion for obtaining ratings of assurance was made by Dr. Anna Shotwell of the Pacific State Hospital staff.

<sup>1</sup> This study was done as part of a project on the selection of Psychiatric Technicians, supported by a grant from the U. S. Public Health Service in contract with the Pacific State Hospital, Spadra, California.



Table 2

Loadings in the Single Common Factor in the Four Raters' Ratings of Effectiveness

Rater	Factor Loading	Communality $h^2$
A	.83	.68
B	.68	.46
C	.16	.03
D	.64	.41

ter-ward transfers. Rater C had very little in common with the other raters. Whether this means that C did not know essentially the same employees as the others or rated them on different qualities we cannot tell from this information alone. Taken at its face value, we might well conclude that C's ratings should receive less weight in a composite, if they were used at all.

The intercorrelations of assurance ratings are shown in Table 3. Since assurance may be assumed to be highly correlated with the degree of acquaintance between rater and ratee, we may conclude from Table 3 that raters B and C had the least in common with respect to ratees whom they knew or did not know. Rater A, who had the greatest communality in her ratings of effectiveness, knew more of the ratees in common with D than with B and C. The factor analysis gave a structure with two common factors.

Taking this information together with that from the analysis of the effectiveness ratings, we conclude that C's lack of communality with the other raters was not due to the fact that he knew different employees. C disagreed with the other raters generally as to relative effectiveness of employees that were

Table 3

Intercorrelations of Ratings of Assurance Connected with the Effectiveness Ratings (N = 716)

	Raters			
	A	B	C	D
A		.23	.22	.50
B	.23		.04	.21
C	.22	.04		.46
D	.50	.21	.46	

Table 4

Rotated Factor Loadings of the Raters with Respect to Their Ratings of Assurance

Rater	Factors		Communality $h^2$
	I	II	
A	.31	.60	.46
B	.06	.33	.11
C	.72	.00	.52
D	.64	.49	.66

rated in common. This disagreement could mean that C emphasized different qualities or it could mean that he rated the same qualities but made different evaluations of employees with respect to those qualities. One might conclude that C's ratings are so inconsistent with those of the consensus that they should not be included in a composite. On the other hand, perhaps C had some neglected valid qualities or some better evaluations to contribute. The best solution seems to be to include C's ratings for what they are apparently worth, that is, to give them a relatively low weight.

The Weighting System

The weighting system we propose and that we have used in connection with each of the Psychiatric Technicians takes into account two variables. One is the factor loading of each rater in the single common factor in the effectiveness ratings. Each rater's ratings, regardless of ratee, is multiplied by this weight. The other weight is the rater's rating of assurance that he applied to each ratee. The over-all weight to be applied to each effectiveness rating is therefore a product of these two values. The composite rating for an employee is a weighted mean of the four effectiveness ratings given him by the four raters.

In order to state more explicitly how the weighted mean is computed we define the following symbols:

Let

- $X_{ik}$  = rating of effectiveness of individual I given by rater K,
- $A_{ik}$  = rating of assurance that rater K makes concerning his rating  $X_{ik}$  of individual I,

$F_k$  = general-factor loading of rater K in his effectiveness ratings,<sup>3</sup>

and

$\bar{X}_i$  = weighted mean of effectiveness rating for individual I.

The equation reads

$$\bar{X}_i = \frac{\sum A_{ik} F_k X_{ik}}{\sum A_{ik} F_k} \quad (1)$$

The summations in both numerator and denominator are over all raters.

### Reliability of the Composites

In order to determine whether the weighting system leads to improvement over a simple summation or average of ratings, we have made a reliability study of composites derived with and without weights. Reliability is defined here as inter-rater consistency or inter-composite consistency. It is impossible to estimate the reliability of composites of all four ratings, but it is possible to estimate reliabilities for composites of two raters at a time. Consequently, the raters were combined in all possible pairs of two and for each pair a weighted and an unweighted composite were computed for each rater. The three possible intercorrelations of such composites, weighted and unweighted, are given in Table 5, based upon 50 randomly selected rates. In every case, the weighted composites show a higher intercorrelation; in two cases very much higher. We have not applied the Spearman-Brown formula to estimate the reliability of composites of four raters for the reason that the conditions for applying that formula are probably not satisfied. The chances are that the reliability of such a weighted composite would be higher than any of the estimates in Table 5. This would indicate that in combining the four ratings for each employee, with the weights, we have a criterion that is sufficiently dependable for use in a valida-

<sup>3</sup> If there should be more than one common factor in such an analysis, the investigator has at least two alternatives. One would be to use the first centroid factor loadings. This would be preferred when other factors are particularly weak. The other alternative would be to use the loadings from each factor (after rotation) separately as a set of weights and to compute a criterion measure corresponding to each factor. Unless these weights were very different, however, the two criteria would be highly correlated.

Table 5

Correlations Between Unweighted and Weighted Composites of Ratings of Effectiveness Assigned by all Possible Pairs of Raters \*

Pairs of Raters	Unweighted Composites	Weighted Composites
AB vs. CD	-.04	.54
AC vs. BD	.58	.64
AD vs. BC	.18	.54

\* From a random sample of 50 rates.

tion study. In view of the two very low correlations for the unweighted composites, there is some question as to whether an unweighted composite of four ratings would be sufficiently dependable to serve as a criterion.

### Summary

This article faces two problems: (1) the fact that different raters in a practical situation do not know employees equally well and thus cannot rate them with equal assurance; and (2) the fact that raters differ with respect to how well they reflect the consensus of the group of raters. The ratings of effectiveness of 716 hospital employees given by four supervisors were studied by factor analysis to determine what their consensus indicated. One common factor, in which raters had quite different factor loadings, was sufficient to account for the intercorrelations of effectiveness ratings. In rating each employee, each rater also gave a rating of degree of his assurance of his correctness.

A factor analysis of these assurance ratings gave two common factors, which were taken to indicate communalities of acquaintance with the employees. The results of the two factor analyses led to the inclusion of the ratings of all raters and to the use of weights in forming composite ratings. One weight was the factor loading of the rater obtained from intercorrelations of the effectiveness ratings. The other weight was the rating of degree of assurance. The composite was a weighted mean of the four ratings of each employee. It was demonstrated that weighted composite ratings based on this principle were definitely more reliable than corresponding unweighted composites.

Received October 11, 1953.

## A Technique for Keying Items of an Inventory to be Added to an Existing Test Battery

Charles O. Neidt

*University of Nebraska*

and

John P. Malloy

*Marquette University*

In developing a quantitative scoring procedure and in selecting items for a test which elicits item responses that cannot be readily classified as correct or incorrect, test constructors have customarily used one or a combination of three procedures. The first procedure is that of having authorities or "juries" select the item response which they believe parallels a definition of the behavior being evaluated. The resulting individual item validity and total test validity are thus dependent upon the judges' interpretation of the defined behavior. The second procedure is that of assigning larger values to those responses internally consistent with the total score. The validity of items keyed and selected according to this procedure depends upon the validity of the total score. The third procedure involves constructing a key and selecting items after correlating each possible item response with an external criterion, usually some behavior display or rating of the subjects. Insofar as subsequent prediction of behavior external to the test score is concerned, the external criterion technique contains inherent advantages. If the criterion against which the items have been validated is a heterogeneous criterion, however, the test will also tend to be heterogeneous. Item selection techniques proposed by Horst (6), Gulliksen (5), Davis (1), and French (4), which combine elements of the second and third procedures, tend to reduce the heterogeneity of the test.

When a battery of tests is used for the prediction of a criterion, maximum predictive effectiveness will occur when each test in the battery has a high correlation with the criterion and a low intercorrelation with the other tests in the battery. Thus if a new

test is to be combined with previously available tests for the prediction of some criterion, then the items in the new test should measure some part of the criterion not already being measured. When individual items of a new instrument are validated against the criterion, the test constructor is usually assured of some subsequent predictive effectiveness when the test is used singly for prediction. If a test validated in such a manner is added to a battery, however, the test constructor has no assurance that the test will increase the total predictive effectiveness of the battery. The reason for this lack of assurance may be that the extent to which the items in the new test intercorrelate with the other tests in the battery has not been considered. The desirability of a technique which takes into consideration that variation in the criterion already associated with other prediction variables is readily apparent.

It was the purpose of this study to determine the relative effectiveness of: (1) keying the items of a new inventory to be added to a test battery in terms of their correlation with the total variation of an external criterion; and (2) keying the same items in terms of their correlation with the criterion variation unexplained by other tests in the battery.

### The Techniques

*The External Criterion Technique.* Because the external criterion technique for keying item responses to attitude, interest, personality, and biographical data instruments has been widely used for the past twenty years, specific instances of its application will not be cited here. Essentially this technique consists of obtaining the correla-



tion between each item response of a key group and a criterion. If a test contains 20 items each having four possible responses, then 80 correlations are obtained. The key is constructed by assigning quantitative values to subsequent responses according to the size and/or direction of the correlations. With such a procedure more than one response can be scored for each item. The total score for subsequently administered tests is then obtained by summing the values assigned to the item responses of each subject according to the key. The desirability of checking the validity of the total score for members of another sample, independent of the key group, should be obvious.

*The Deviate Technique.* The procedure of keying item responses according to their correlation with the unexplained criterion variation, here referred to as the deviate technique, is much less well known than the external criterion technique. Instances in which the deviate technique has been used include the research of Neidt and Merrill (9) and Neidt and Edmison (10). In an article published in 1951, Meyers and Schultz (8) described a modified version of this technique, and in an article appearing in 1953, Schultz and Green (11) reported the use of the deviate technique in a way similar to that used in the present study.

In constructing a key with the use of the deviate technique, the responses of a key group to each item are correlated with that part of the criterion variation which is not associated with other test scores in a battery. In the analysis of regression of a test battery and a criterion, the criterion variance unexplained by other tests can be expressed for any group as follows:

$$\Sigma y^2 - [a_1 \Sigma x_{1y} + \dots + \Sigma a_m \Sigma x_{my}]$$

where  $\Sigma y^2$  is the criterion sum of squares, the  $\Sigma x_{iy}$ 's are the sums of the cross products of the test scores and the criterion in deviation form and the  $a$ 's are regression weights determined by least squares. The foregoing expression can be readily changed to raw score form. For any individual in the group for which the regression weights have been de-

termined, an indication of the unexplained variation may be obtained from  $V - \hat{V}$ , in which  $V$  is the actually obtained criterion measure, and  $\hat{V}$  is the criterion measure predicted for this individual from scores in the test battery. After prediction and subtraction from the actual criterion measures have been made for each individual in a key group, a distribution can be formed which represents that variation in the criterion that is unaccounted for by the tests in the battery. This distribution will be distributed around zero and its shape, although influenced by the shape of the criterion distribution, will tend toward normality. It is this distribution of actual-minus-predicted criterion measures with which item responses are correlated in the use of the deviate technique.

### Procedure

*Collection of Data.* A 201-item life experience and attitude toward education inventory, constructed by Malloy (7), was administered to 309 freshman women entering the University of Nebraska in September, 1952. Of the 201 items in the inventory, 112 were of the multiple-choice type and 89 were of the paired-statement type. The items were designed to reflect experiences and attitudes in four areas, viz., school experiences and attitudes toward education, self appraisal, family relationships, and choice of friends.

The 309 students were subdivided into two independently drawn random subsamples of 155 and 154 students each. The sample containing 155 students was designated as the key group and the sample of 154 was designated as the cross validation group.

Since the inventory was constructed to be used with a battery of two other preregistration tests, scores on these tests were obtained for both groups. The two preregistration tests involved were the American Council on Education Psychological Examination, Linguistic subtest, and a local English achievement test, entitled the English Placement test. Raw scores are customarily converted to a one-to-nine scale at the University of Nebraska and these converted scores were used in this investigation.

The criterion used in this study was first-semester average course mark. Course marks are also reported on a one-to-nine scale at the University of Nebraska, nine being the highest mark and one signifying failure. Weighted averages for the students were obtained according to the hours of credit involved for individual courses.

Thus the data for this study, other than scores on the inventory, included: first-semester average

Table 1

Weights Assigned to the Item Correlations  
for Each Key

Correlation	Weight
0.25 or higher	+2
0.10 to 0.24	+1
-0.09 to 0.09	0
-0.10 to 0.24	-1
-0.25 or higher	-2

course marks, ACE-L scores, and English Placement scores for two groups of 155 and 154 students each.

*Development of the Keys.* To develop the two keys for the inventory, two separate analyses were made of each item response. In constructing the key according to the external criterion technique, the correlation between each item response and the criterion was estimated with the use of Flanagan's correlation table (3). In constructing the key according to the deviate technique the correlation between each item response and the distribution of actual-minus-predicted course marks was obtained in the same manner. The regression equation used in obtaining the actual-minus-predicted distribution was

$$\hat{Y} = .197 X_1 + .195 X_2 + 4.125$$

where  $X_1$  is English Placement score in stanine form and  $X_2$  is ACE-L score in stanine form.

After the two complete sets of correlations had been obtained, the two keys were constructed by assigning weights to each item response according to the size and direction of the correlation with each criterion. The weighting system used for each key is shown in Table 1.

The limits of the intervals in the distribution of correlations as well as the weights were arbitrarily designated. Because the use of Flanagan's table for estimating correlations involves

only the upper and lower 27 per cent of the criterion distribution, the significance from zero of the estimated correlation coefficients was not ascertained.

The degree of similarity between the weights assigned to each item response for the two keys may be seen from Table 2. Each of the 201 items of the inventory contained from two to five response choices which yielded the total of 629. The coefficient of correlation between the two distributions of response weights shown in Table 2 is .509. The deviate technique key contained 368 item response weights other than zero as compared with 339 such response weights for the external criterion technique key. It should be recalled that in responding to the inventory, however, each subject gave 201 responses, rather than 629.

The number of items having response weights of zero for all choices within the item was found to be 53 for the external criterion technique key and 42 for the deviate technique key. Of the items having all response weights of zero, 34 such items appeared in both keys. In summary, eleven more items in the deviate technique key than in the external criterion technique key contained one or more response weights other than zero.

The inventories of the 154 students in the cross validation group were scored using each of the two keys. To avoid negative scores, the constant 50 was added to each of the two inventory scores for the subjects in the cross validation group. The correlations between each of the two inventory scores and the criterion and between these scores and the other test scores in the battery were computed. The significance of the contribution of each independent variable to the prediction scheme was ascertained by analysis of regression.

Results

In Table 3 are shown the zero order coefficients of correlation between the variables in this study. It is interesting to note that

Table 2  
Item Responses Classified by Weight According to Two Keys

External Criterion Key Weight	Deviate Technique Key Weight					Total
	-2	-1	0	+1	+2	
+2	0	3	7	12	16	38
+1	1	19	43	49	9	121
0	10	51	162	55	12	290
-1	12	57	43	17	0	129
-2	21	21	6	2	1	51
Total	44	151	261	135	38	629



Table 3

Zero Order Coefficients of Correlation Between Each  
Pair of Variables for 154 Cross  
Validation Students

	$X_1$	$X_2$	$X_3$	$X_4$
$Y$	.446	.446	.512	.332
$X_1$		.735	.480	.478
$X_2$			.334	.304
$X_3$				.624

$Y$  = Average Course Mark.

$X_1$  = External Criterion Technique score.

$X_2$  = Deviate Technique score.

$X_3$  = English Placement score.

$X_4$  = ACE-L score.

the two inventory scores yielded correlations of the same magnitude with the criterion. It should also be noted that, in general, the deviate technique score correlated lower with the other scores in the battery than the external criterion technique.

In Table 4 are shown the multiple and partial correlation coefficients of the combined variables. Inspection of Table 4 indicates that the deviate technique score contributed significantly to the effectiveness of the total battery, whereas the external criterion technique score did not. In addition, the optimal combination of two prediction variables in-

Table 4

Multiple and Partial Correlation Coefficients  
for Combined Variables

Multiple Correlations	Partial Correlations
$R_{Y(X_1X_2X_3X_4)} = .589$	$r_{YX_2 \cdot X_1X_3X_4} = .329^{**}$
$R_{Y(X_2X_3X_4)} = .587$	$r_{YX_1 \cdot X_2X_3X_4} = .055$
$R_{Y(X_1X_3X_4)} = .517$	$r_{YX_1 \cdot X_2X_4} = .084$
$R_{Y(X_1X_2)} = .516$	$r_{YX_2 \cdot X_3X_4} = .358^{**}$
$R_{Y(X_2X_3)} = .570$	
$R_{Y(X_1X_4)} = .362$	
$R_{Y(X_2X_4)} = .485$	
$R_{Y(X_3X_4)} = .515$	

$Y$  = Average Course Mark.

$X_1$  = External Criterion Technique score.

$X_2$  = Deviate Technique score.

$X_3$  = English Placement score.

$X_4$  = ACE-L score.

\*\* Indicates a partial correlation coefficient significantly different from zero at the 1 per cent level.

cludes the English Placement test score and the deviate technique score of the inventory.

### Discussion

The empirical results from this investigation indicate that the deviate technique is superior to the external criterion technique for keying items of a new test to be added to an already existing battery. If the key for the life experiences inventory used in this investigation had been constructed using only the external criterion technique, the inventory would not have significantly increased the predictive effectiveness of the battery.

The similarity of the zero order correlations, .446 and .446, between average course marks of the cross validation group and the two inventory scores is striking. It is doubtful that such close correspondence will be found in subsequent studies of a similar nature. In general, it seems reasonable to postulate that the more homogeneous the criterion, the more divergent such coefficients of predictive effectiveness for the two techniques will become, i.e., with a homogeneous criterion the external criterion technique correlation will probably be higher than that found for the deviate technique. The similarity of the two coefficients found in this study is perhaps the result of the heterogeneity of the criterion.

The assignment of weights ranging from  $-2$  to  $+2$  to the item responses of the inventory imposed a condition of item selection on the keying procedures. Some items were assigned to weight of zero according to one keying technique and weights other than zero according to the other technique. Such differences between the weights assigned to the item responses will influence the apparent length of an instrument and the variability among the resulting total scores. Thus in comparing validity coefficients to evaluate two keying techniques, consideration should be given to differences between measures of central tendency and variability of the total score distributions. If the variability of the distribution obtained by one keying technique is considerably larger than the variability of the other distribution, differences between validity coefficients could result



which are attributable to differences between the total score variabilities rather than to actual differences in effectiveness of the techniques. When the means and standard deviations for the two total score distributions involved in this study were computed, the means were found to be 55.44 and 54.06 and the standard deviations were found to be 17.59 and 16.02 for the external criterion technique and the deviate technique, respectively. Because these differences are so small, it is felt that the greater contribution made by the deviate technique key scores to the predictive effectiveness of the total battery was not attributable to the item selection imposed by the weighting procedure. Apparently the scored items of the deviate technique key contained more similar response weights within the items than the scored items of the external criterion technique key. Such a condition could result in a larger mean and standard deviation for the external criterion technique key total scores.

The fact that the score on the life experience inventory contributed significantly to the prediction of average course marks suggests the importance of evaluating other characteristics of students than scholastic aptitude and achievement. A detailed description of the content, construction, and analysis of the instrument used as a vehicle for this study will be published subsequently.

### Summary

The purpose of this study was to determine the relative effectiveness of keying the items of an inventory to be added to an already existing test battery according to: (1) the correlation of the item responses with the total variation in a criterion (first semester average course marks); and (2) the correlation of the same item responses with the criterion variation unexplained by other tests in the battery. Two sets of keys were constructed based upon the responses of 155 subjects. Each inventory of 154 subjects constituting a cross validation group was then scored using the two keys. The zero order correlations between the score derived from each key and the criterion were found to be

identical for the 154 subjects in the cross validation group. When the two scores were combined with others in a test battery the contribution to the predictive effectiveness of the total battery made by the key derived from correlating item responses with the unexplained variation was found to be significant. The contribution made by the key derived from correlating item responses with the total criterion variation was found to be not significant.

Received October 21, 1953.

### References

1. Davis, F. B. Item selection techniques. In E. F. Lindquist (Ed.), *Educational measurement*. Washington: American Council on Education, 1951.
2. Davis, F. B. Item analysis in relation to educational and psychological testing. *Psychol. Bull.*, 1952, 49, 97-121.
3. Flanagan, J. C. *A table of values of the product-moment coefficient of correlation in a normal bivariate population corresponding to given proportions of successes*. New York: Cooperative Test Service, 1936.
4. French, J. W. A technique for criterion-keying and selecting test items. *Psychometrika*, 1952, 17, 101-106.
5. Gulliksen, H. O. *Theory of mental tests*. New York: Wiley, 1950.
6. Horst, A. P. Item selection by means of a maximizing function. *Psychometrika*, 1936, 1, 229-244.
7. Malloy, J. P. *The prediction of college achievement with the life experience inventory*. Doctoral Dissertation, University of Nebraska, 1953.
8. Meyers, R. C. and Schultz, D. G. Predicting academic achievement with the use of a new attitude interest questionnaire, I. *Educ. psychol. Measmt.*, 1950, 10, 654-663.
9. Neidt, C. O. and Edmison, L. D. Qualification responses used with paired statements to measure attitudes toward education. *J. educ. Psychol.*, 1953, 44, 305-311.
10. Neidt, C. O. and Merrill, W. R. Relative effectiveness of two types of response to items of a scale on attitudes toward education. *J. educ. Psychol.*, 1951, 42, 432-436.
11. Schultz, D. G. and Green, B. F., Jr. Predicting academic achievement with the use of a new attitude-interest questionnaire, II. *Educ. psychol. Measmt.*, 1953, 13, 54-64.
12. Super, D. E. The validity of standard and custom-built personality inventories in a pilot selection program. *Educ. psychol. Measmt.*, 1947, 7, 735-744.

# The Effect of Age and Experience upon Accident Rate

R. H. Van Zelst

*Kroh-Wagner Co., Riverside, Illinois*

Industrial accidents, their prediction and control and the various factors related to and affecting them have long been a subject of study for the psychologist in industry. One of the specific topics of interest has been the relationships existing between the age and experience of the worker and his accident frequency.

Most research studies in this area have demonstrated the existence of some relationship between accidents and both experience and age. Though by no means universal the general conclusion arrived at in these experiments is that accident frequency tends to decline with increasing age and/or experience.

Many of the studies of experience suffer, however, from a procedural error. The most common method applied in this type of study appears to be to divide the men in a given organization into experience groups and then to calculate the accident rate of each group. The application of this method of necessity assumes that if no differences in experience exist, all of these different groups would have the same average number of accidents. However, it is also reasonable to assume that in many jobs the high-accident employees will tend to drop out either through retirement due to injury, separation or voluntarily leaving employment. Such a natural selection process tends to retain on the job only those persons who have maintained a certain safety standard in their operations.

The usually discovered decrease in accident frequency with experience may be due then to this natural selection process. What would then appear to be necessary in order that the effects of experience may be properly evaluated is to follow the accident history of the same group of workers over a period of time. Several studies (1, 2, 3, 5) have done this. Unfortunately, in most instances these studies either follow the employee's accident history for only a relatively short duration or

fail to remove possible influences due to the operation of the age variable.

The study of the relationship between age and accident frequency presents a somewhat similar picture to the experience problem. The typical procedure here again is to subdivide employees into differing age groups and to compute the mean number of accidents for each age group. In most instances, however, age is highly correlated with experience, thus confusing the issue and making it difficult, if not impossible, clearly to ascribe any discovered relationship either to age or to experience.

Attempts have been made to minimize the effect of the experience variable through the utilization of partial correlational methods (1, 3). However, these methods are also subject to question in that it is not certain that experience may be held constant by using partial correlation methodology in view of the safety selection process previously mentioned. It seems probable that the operation of these selective factors prevents compliance with certain basic assumptions inherent in this statistical method.

It is the author's purpose therefore to present material obtained in a different manner from most of the previous studies in this field in an attempt to provide more information and gain further insight into the existence of the relationships between age and experience with accident frequency.

## Subjects

The subjects used in this study are employees of a copper plant in Indiana. These subjects were selected from six sections comprising a single large department operating metal forming mills. Work tasks were identical for the members of all groups and no unusual differences in pressures of production were observed for the different groups during the periods of data collection.

Conditions of work, light, heat, ventilation

were also highly similar for all subjects as were the number of hours worked. Only employees working on the same shift were used in this experiment.

Conditions and methods of work, together with type of equipment, remained virtually constant throughout the five-year experimental period.

A total of 1,237 employees who remained with the company in the above mentioned department for the experimental period had their accident records carefully traced and charted for each month of the period. In addition other members of the work force hired at the same time (when the plant was first opened) but who dropped out or were separated also had their records carefully tabulated and recorded. These workers at the onset of the experimental period totaled an additional 1,317 workers.

The number of accidents experienced by each man was readily traced through employee history data which contained carefully detailed records of dispensary visits and their reason and cause. It is felt that this criterion is valid since it is a compulsory policy of the company to have all employees who are injured on the job, regardless of how slight the injury might be, visit the dispensary for medical clearance, treatment, and report. No distinctions as to severity of accident were made in this study. Only accidents occurring during working hours and in actual performance of the job were used.

Accident frequency data were reported on the basis of mean number of accidents per 1,000 man hours of operation. Payroll records of the subjects provided the necessary data for computation.

### Results

Figure 1 displays graphically the mean number of accidents per 1,000 man hours of operation for both of the experimental groups and also the entire departmental mean accident rate. Accident rate figures are reported on a monthly basis for a period of 60 months or 5 years. Accident rates for the turnover group are not reported after the first 30 months because of the small number of workers remaining in that group beyond this

period of time. (The number of workers in the turnover group was reduced to 243 members at the end of thirty months.)

It can readily be seen from the presented data that in this particular instance the accident rate for these workers declines rapidly during the first five months of operation for both of the groups. The entire department mean accident rate closely approximates the rate curves of the two experimental groups. This is readily explained by the fact that the two experimental groups, particularly in the early phases of the experimental period, comprised a majority of the entire work force.

The tendency for the departmental rate curve to be higher during the latter phases of the experimental period can be attributed to the incorporation of newly hired employees into the work force.

The consistency with which the accident rate curve of the turnover group remains higher than that of their fellow-workers tends to support the hypothesis that a natural selection process does exist. The higher rate and more gradual decline in accident frequency for this group of turnover employees apparently is indicative of an informal and perhaps to some extent a formal weeding out of high-accident workers.

In studying these accident rate graphs the effect of job experience upon the accident rate of these workers appears to be considerable for their first five months of employment, but seems to be of little significance beyond the fifth month of employment. The general leveling off in accident rate after five months on the job seems to point up the thesis that experience makes its contribution towards accident rate reduction by familiarizing the employee with proper work and safety habits. Apparently five months of on-the-job duties is sufficient for these workers on this particular type of operation to become well enough trained to reduce accident rate to what may be considered normal expectancy. It should be pointed out that these workers did not receive the benefit of any formalized pre-job assignment training and so actual experience was called upon to substitute for this formalized training.

These initially high accident rates would



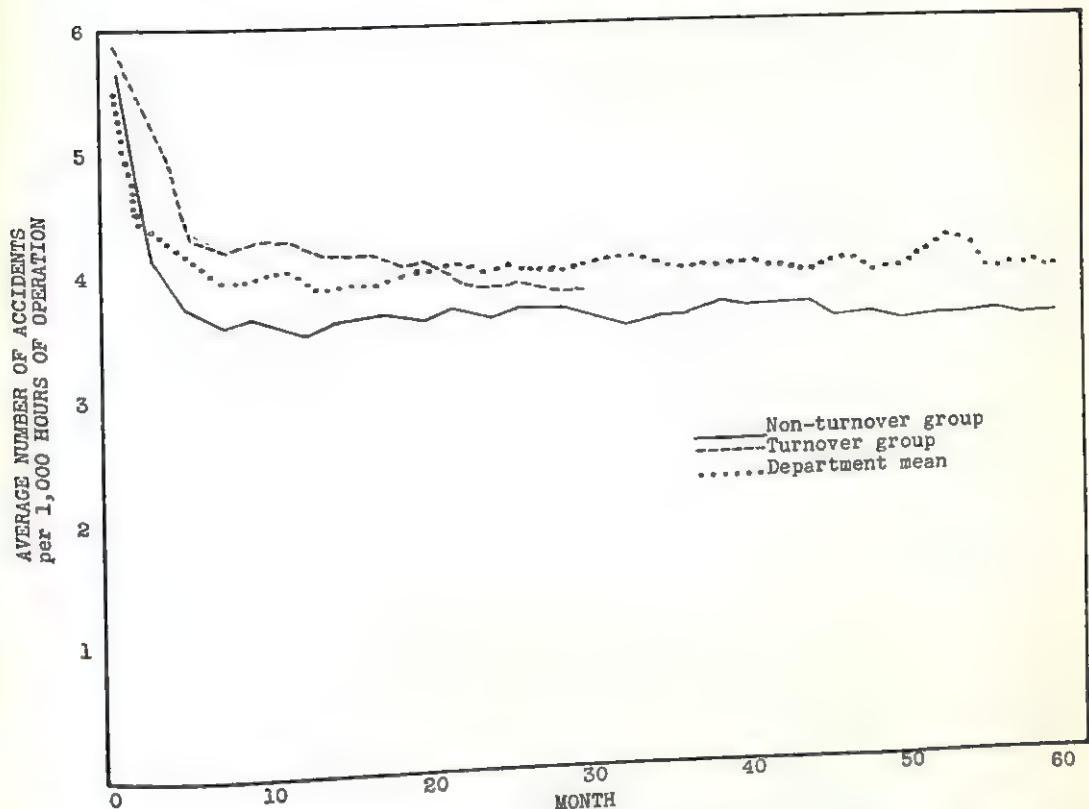


FIG. 1. The relationship between experience on the job and the average monthly accident rate per 1,000 hours of operation for a non-turnover group and a turnover group.

appear to lend further support to the often stressed necessity for proper immediate training in correct work methodology and safety habits.

In order to provide an experimental test of this conclusion the accident rates of another 872 workers were charted. These men had been hired at various times after the company was better established and so received the benefit of formal training in correct job procedure and safety methods. These men also performed the same work tasks under highly similar conditions. Data on this group for their first fifteen months of employment are graphically presented in Figure 2.

Results here follow the same general pattern found for the previous groups. There is an almost identical sharp decline in accident frequency for the early on-the-job period followed by the same leveling off pattern. Of note, however, is the fact that the initial accident frequency rate is markedly lower for

this group. Furthermore, the level which approximates what has been termed normal expectancy for the previous groups is reached after the third month of on-the-job performance rather than after the fifth.

In view of the strong similarity between the work tasks and work environment of this and the other two previous groups, this reduction in the frequency of accidents amongst these workers for this formative period can in the author's opinion be traced only to the benefits derived from the formal training program.

However, the observed decline, still sharp, for these trained workers during the early phases of their employment still suggests the importance of actual accumulated on-the-job experience in bringing accident rates down to what might be considered normal.

Still untested is the effect of age upon accident frequency. To study this relationship two other groups were formed. These groups

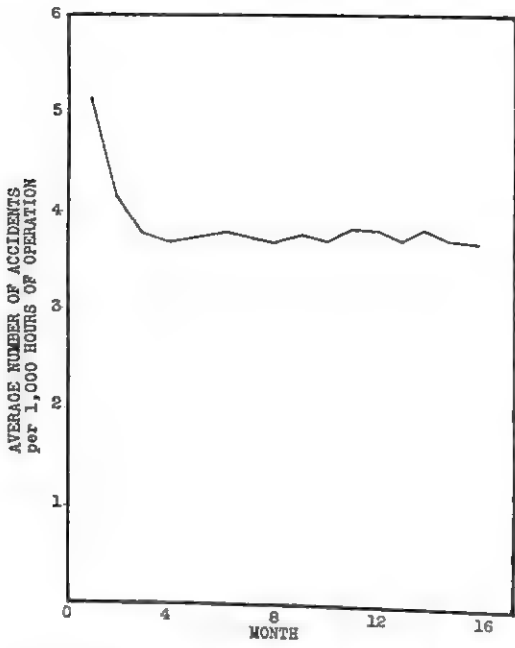


FIG. 2. The relationship between experience on the job and the average monthly accident rate per 1,000 hours of operation for a group of trainees.

were matched on the experience variable. Group A was a young group (Mean Age = 28.7 years, S.D. = 1.4, N = 639) with approximately three years of experience (Mean experience = 2.9 years, S.D. = .45). Group B was composed of older workers (Mean Age = 41.1, S.D. = 2.9, N = 552) also with approximately three years of experience (Mean experience = 3.2 years, S.D. = .63).

Accident frequency rate for these groups (Figure 3) differs markedly throughout the eighteen month experimental period. Although both groups have the same amount of experience, the younger group has what appears to be a significantly higher accident rate than their older work companions.

As might be expected the younger group's (Group A) accident rate is above the department level while the mean accident rate of the older group (Group B) is below the department's level for these particular periods of time.

To further pursue this study of the effects of age upon accident frequency rate a third group (Group C) was used. These workers were similar to Group B in that they too were an older group (Mean Age = 39.2, S.D.

= 3.1, N = 297), but unlike either of the two previous experimental groups these men were inexperienced at the onset of the experimental period. They did, however, receive the benefit of training prior to actual job assignment and performance.

As can be seen from Figure 3, the accident frequency rate for this group again as in past instances shows the same early sharp decline followed by a general leveling off to a position approximating that of the older group (Group B). The accident rate of Group C follows also the pattern of the previous trainee group although mean accident frequency is somewhat lower throughout the period.

It is to be noted that from the third month onward and practically from the second month onward the accident rate for this group of workers is lower than that of their younger and much more experienced fellow workers (Group A). It is also to be noted that this older group functions below the mean departmental level after what might be termed the three-month breaking-in period.

The greater strength of the relationship between age and accident frequency rate as compared with experience and accident frequency rate becomes even more noticeable as

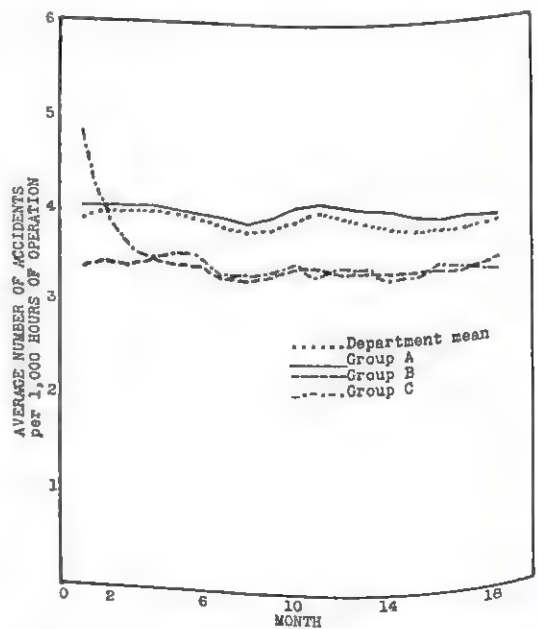


FIG. 3. The relationship between the age of the employee and average monthly accident rate per 1,000 hours of operation.

the experience level differences begin to disappear with the passage of time spent on the job.

It would appear then from these data that age is definitely related to accident frequency. Older employees in this study even when less experienced maintain better safety records than do the younger men. The accident rate of these younger workers exceeds slightly the mean accident rate level of the entire department despite the disparity in job experience. Their rate, in fact, appears from the data to be exceeded only by those employees who are currently in the breaking-in stage of development.

### Summary and Conclusions

The results obtained in this experiment seem to indicate that at least for these groups of men and for this particular type of operation the effect of experience upon the frequency rate of accidents apparently is limited to a three to five month period of initial on-the-job performance. This particular period of time may be termed a breaking-in period and it is characterized by a sharp decline in the number of accidents. Following this period there is a leveling off in accident rate throughout the employee's work history. This rather level period may be considered to be normal expectancy.

When the workers are given formal training prior to actual job performance there is a considerable reduction in early accident frequency rate, which is manifested in lower initial accident frequency and also in what may be regarded as a faster developmental period in that the amount of time required for the trained work groups to level off at the normally expected frequency is significantly reduced.

It would appear that age in this instance apparently exerts a greater influence upon accident rate than does experience once the breaking-in stage is passed. From the comparisons made between the matched work groups it has been found that older workers tend to have fewer accidents than their

younger co-workers. This appears to be true throughout the employee's work history when similar groups are compared. Lower accident rates are remarkably characteristic of these older men from their earliest job performance on.

It is the author's opinion, although no conclusive evidence is presented, that since age exerts the stronger influence upon accident frequency rate, beyond initial employment, it is necessary to explain accidents in part on the basis of immaturity of employees. Furthermore, the usually found reduction in accident rate with increasing age and experience can also be attributed to some extent to the operation of a natural selection process which results in the weeding out of workers less fit for the job. It is also felt that little importance can be attached to the effect of experience upon accident rate for periods other than that of initial employment particularly when the effects of age and the natural selection process are eliminated. Proper training in correct work methodology and safety habits can further reduce the effect of experience upon accident rate but cannot apparently substitute completely for actual job performance in helping the worker to internalize fully the correct procedures and habits necessary to efficient operation from the safety standpoint.

Received September 28, 1953.

### References

1. Brown, C. W., Ghiselli, E. E. and Minium, E. W. *Experience and age in relation to proficiency of street car motormen*. Report to Municipal Railway System of San Francisco, 1946.
2. Chaney, L. W., and Hanna, H. S. *Safety movement in the iron and steel industry*. Bur. Labor Statistics, Rept. 234, 1918.
3. Hewes, A. Study of accident records in a textile mill. *J. Industrial Hygiene*, 1921, 3, 187.
4. Newbold, E. M. *A contribution to the study of the human factor in the causation of accidents*. Ind. Fatigue Research Bd., Rept. 34, 1926.
5. Vernon, H. M. Prevention of accidents. *Brit. J. ind. Med.*, 1945, 2, 3.



## Note on Age and Productive Scholarship of a University Faculty

Robert A. Davis

George Peabody College for Teachers

The results presented are a part of a larger study conducted for the Council on Research and Creative Work of the University of Colorado in 1946. The study was designed to survey the research and writing activity of the entire faculty (representing all the schools and colleges) during a twenty-year period, 1920-1939 inclusive. This is a period that we believed would reflect trends between two major wars—a relatively stable period in the history of the university.

During the period covered by the study faculty members had been requested annually to submit to the Dean of the Graduate School a list of papers, articles, and books written during the year just ended; and these items were published annually in the Graduate Bulletin. In order to safeguard accuracy the author sent each faculty member a list of his contributions as recorded in the Graduate Bulletin and requested that they be checked.

The terms *research* and *writing* should be noted. No effort was made to differentiate between items that were definitely of research character and those that were scarcely more than descriptive or expository documents. Also attention is called to the term *activity*. The study did not deal with the difficult problem of appraising contributions of faculty members. Instead, it was concerned exclusively with the amount of research and writing completed.

The data reported here concern only one aspect of the larger study, that of research and writing in relationship to the age of the faculty member at the time. During the period covered by the study any person contributing one item was regarded as writing. Co-authors were treated in the same manner as authors writing independently. In cases of multiple authorship each person received the same credit that he would have received

as a single contributor. The curves show *absolute* and not *proportionate* numbers of contributions. Consequently, they do not make allowances for the diminishing numbers of potential contributors at the upper age levels. Figure 1, which is based on the records of 385 faculty members, tells the story.

The results suggest a number of questions. How do research and writing activity relate

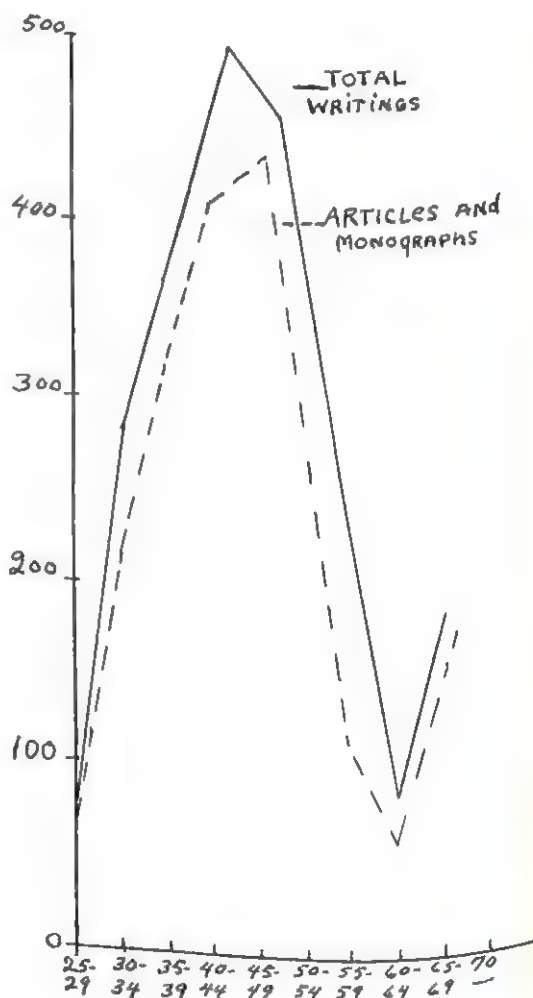


FIG. 1. Number of items published and age of contributor.

to the age at which a faculty member attains full professorial status? How do they relate to salary increases and promotional policy in general? What should be the policy of a university administration regarding research and writing? What means may be used to stimulate research? Other kinds of professional growth? If faculty members as a group reach a peak in research and writing

activity around 45 years of age is there evidence that they continue to grow professionally in other respects? Is there any fundamental reason for the peak of activity around 45 years of age? Is this a crucial period in the career of a faculty member? The reader will think of many other questions.

*Received October 16, 1953.*

## Relationship of Employee Morale to Ability to Predict Responses<sup>1</sup>

Rossall J. Johnson

*School of Commerce, Northwestern University<sup>2</sup>*

This investigation is concerned with the relationship between the morale of an employee and the ability of the employee to predict the responses of his subordinates and the morale of these subordinates.

There has been some evidence (1, 4, 5) to indicate that where individuals "knew" and understood one another they were able to predict the others' responses. It would seem to follow from this that where a group and leader relationship existed, the ability of the group members to predict the leaders' responses would be dependent upon how well these group members understood their leader. And conversely, the ability of the leader to predict the group members' responses would be dependent upon how well the leader understood the group members.

This problem may be clarified by asking three questions. 1. Do subordinates with high morale "know" and understand their supervisor better than low morale subordinates? 2. Is the morale of the subordinates who "know" and understand their boss higher than those who do not "know" and understand their boss? 3. Does the supervisor "know" and understand the high morale subordinates better than the low morale subordinates? If the answers to questions 1 and 2 are yes, then one may anticipate the development of a questionnaire which will indicate morale by measuring the ability of the subordinate to predict the responses of his supervisor. An affirmative answer to question 3 would indicate that the morale of the subordinates may be estimated by measuring the ability of the supervisor to predict the responses of his subordinates.

In order to analyze this problem, the following null hypotheses were set up:

1. There is no significant difference between the morale scores of subordinates who can predict the responses of their supervisors best, and the morale scores of subordinates who have the least success in predicting the responses of their supervisors.

2. There is no significant difference between the ability of high morale subordinates to predict the responses of their supervisors and the ability of low morale subordinates to predict the responses of their supervisors.

3. There is no significant difference between the morale scores of individual subordinates whose responses were most successfully predicted by their supervisors and the morale scores of individual subordinates whose responses were least successfully predicted by their supervisors.

### Procedure

A sample of 227 subordinates and 25 supervisors was taken from two companies. The subordinate, for the purpose of this study, is designated as a randomly selected hourly paid worker who does not have group leader responsibility and who has worked for the tested supervisor for at least nine months. The supervisor is defined as a salaried supervisor who has at least 12 subordinates (as defined above) reporting directly to him. This supervisor should have supervised these 12 subordinates at least nine months. Eight to 10 subordinates under each of the 25 supervisors participated in the project.

Three scores were calculated from the questionnaire: (1) subordinate morale score; (2) supervisor predicting score; and (3) subordinate predicting score. The subordinate morale score is the number of times, out of a possible 20, that the subordinate selected the most favorable response. The supervisor predicting score is the total number of times the supervisor correctly predicts the subordinate's response to 20 questions.

The subordinate questionnaire consisted of three parts. Part A was a selection of 20 questions from form A of the test *How Supervise?* These questions were from the sections on supervisory practices and supervisory opinions. As shown in the example, the question mark or undecided response alternative was omitted. Part

<sup>1</sup> This paper was presented at the MPA annual meeting, Columbus, Ohio, April 30, 1954.

<sup>2</sup> This is part of a doctoral dissertation done under the direction of Dr. H. H. Remmers of Purdue University.



B consisted of 20 morale questions. In a previous study (2) these questions had D values (3) of 1.10 or higher. Part C consisted of the same questions as in part A but with instructions to predict the response that the subordinate thought his supervisor would give to each question.

The subordinate was guaranteed anonymity. His name and personal data were on a separate sheet deposited in a ballot type of box while the questionnaire with the supervisor's name only was deposited in another box. This questionnaire and the personal data sheet were later brought back together by means of a code.

The supervisor questionnaire consisted of two parts. Part A is made up of the same 20 questions from form A of *How Supervise?* as were used in the subordinates' questionnaire. The supervisor answered these questions as he would if he were answering the complete form, except that the question mark or undecided response alternative was omitted. Part B also consisted of the same 20 *How Supervise?* questions and 20 morale questions mentioned above. With part B the supervisor was given a list of names of the subordinates who filled in the subordinate questionnaire. A maximum of ten names was on the list. Each man had a code number assigned, such as John Jones—No. 1, Bill Smith—No. 2, Ted Green—No. 3, etc. The supervisor was asked to predict how each subordinate answered the 40 questions. For example: when the foreman had decided how John Jones had answered a question, he wrote the number "1" after the predicted answer. He then predicted the responses of subordinates Bill Smith, Ted Green, and so on in the same manner until he had predicted the eight or ten subordinates' responses to all of the questions.

### Results

Supervisors predicted the responses of 25% of the subordinates with scores of 14 or higher. This constituted the high group. Supervisors predicted the responses of 23% of the subordinates with scores of 10 or lower. This made up the low group. A t test was made to determine if there was a significant difference between the mean morale score of the high group and the mean morale score of the low group.

As shown in Table 1, the hypothesis that there is no significant difference between the morale scores of individual subordinates whose responses were most successfully predicted by their supervisors and the morale scores of individual subordinates whose responses were least successfully predicted by their supervisors was not rejected. The t

Table 1

Comparison of Mean Subordinate Morale Scores for High-Low Supervisor Predicting Scores

	Supervisor Predicting Scores	
	High 25%	Low 23%
Mean Subordinate Morale Score	12.3	11.8
t value	.6	
Significance Level	—	

value indicates that there evidently is no significant difference between the means of the individual morale scores.

Nineteen per cent of the subordinates had scores of 17 or higher on the morale survey questions. This group was considered the high morale group. Twenty per cent of the subordinates had scores of 8 or lower. This group constituted the low morale group. The subordinate individual predicting scores on the 20 questions of form A of *How Supervise?* for the high morale group were added and a mean score obtained.

The mean score for the low morale group was obtained the same way. A t test was made to see if there was a significant difference between the mean of the high morale group and the mean of the low morale group.

Table 2 shows that the hypothesis that there is no significant difference between the ability of high morale subordinates to predict the responses of their supervisors and the ability of low morale subordinates to predict the responses of their supervisors was rejected. The t value indicates that the high morale subordinates' mean score was sig-

Table 2

Comparison of Mean Subordinate Predicting Scores for High-Low Subordinate Morale Groups

	Morale Scores	
	High 19%	Low 20%
Mean Subordinate Predicting Score	12.6	9.8
t value	13.8	
Significance Level	1%	

nificantly higher than the low morale subordinates'. The difference in the means was significant at the 1% level.

The mean of the subordinates' morale scores for the subordinates who were most successful in predicting their supervisor's responses was tested by the *t* test method to see if it was significantly different from the mean of the subordinates' morale score for the subordinates who had the least success in predicting their supervisor's responses. Subordinates who predicted the responses of their supervisors most successfully were those with prediction scores of 14 or higher. This group represented the top 24%. Subordinates who predicted the responses of their supervisors least successfully were those with predicting scores of 9 or less. This group constituted the bottom 20%. The mean of the subordinate individual morale score for the subordinates who were most successful in predicting their supervisor's responses was tested by the *t* method to see if it was significantly different from the mean of the subordinates' individual morale score for the subordinates who had the least success in predicting their supervisor's responses.

The hypothesis that there is no significant difference between the morale scores of subordinates who can predict the responses of their supervisors best and the morale scores of subordinates who have the least success in predicting the responses of their supervisors was rejected. Table 3 shows that the mean morale score of the subordinates who had high individual predicting scores was significantly higher than the mean morale score of low predicting subordinates. This differ-

ence as shown by the *t* value was significant at the 1% level.

Discussion and Conclusions

Based on these data, the following conclusions may be drawn.

1. It cannot be generalized as to the morale state of the subordinate and the ability of the supervisor to predict his response. The supervisor is evidently able to predict the responses of some low morale subordinates with as much skill as some high morale subordinates. The non-rejection of hypothesis 1 might be explained by the fact that some low morale subordinates vociferate their objections or criticisms of certain things. Because they expressed themselves forcefully, the supervisor remembers these comments and is thus in a better position to predict the low morale subordinate responses than he is able to predict the responses of those who have average morale.

2. High morale subordinates are better predictors of their supervisors' responses than are low morale subordinates.

3. The subordinates who could predict the responses of their supervisors best had higher morale than those subordinates who had the least success in predicting their supervisors' responses.

These last two conclusions may be interpreted as meaning that those who were better acquainted with their supervisor had higher morale and those who had high morale "knew" and understood their supervisor better.

In connection with conclusion No. 2 there was a possibility that high morale employees were assigning more of their own responses to the supervisor than were the low morale employees. If projection were taking place more with high morale employees than with the low morale employees, the difference in their ability to predict responses might be more a measurement of differences in projection.

To investigate this phase, the answers marked by the employee on the *How Super-wise?* test were compared with the answers the employee predicted his supervisor made. The number of times the answer and pre-

Table 3

Comparison of Mean Subordinate Morale Scores for High-Low Subordinate Predicting Scores

	Subordinate Predicting Scores	
	High 24%	Low 20%
Mean Morale Score	13.4	10.2
<i>t</i> value	3.64	
Significance Level	1%	

diction differed were tallied. It was found by the t test method that the average number of responses which were different for the high morale group were not significantly different from the average number of answers which were different for the low morale group.

#### Summary

An analysis was made to see if supervisors are able to predict the responses of high morale subordinates more successfully than those of low morale subordinates. An analysis was also made to see if there was a difference in morale scores of those subordinates who were able to predict their supervisors' responses most successfully and those who were least successful in predicting the supervisors' responses. The results indicate that supervisors are not able to predict the responses of high morale subordinates with any more success than the responses of low morale subordinates. The results also indicate that

high morale subordinates are able to predict their supervisors' responses better than low morale subordinates.

*Received July 2, 1954.*

*Early publication.*

#### References

1. Dymond, R. F. Personality and empathy. *J. consult. Psychol.*, 1950, 14, 343-350.
2. Harris, F. J. The quantification of an industrial employee survey. I. Method. *J. appl. Psychol.*, 1949, 33, 103-111.
3. Lawshe, C. H., Jr. A nomograph for estimating the validity of test items. *J. appl. Psychol.*, 1942, 26, 846-849.
4. Miller, F. G. and Remmers, H. H. Studies in industrial empathy. II: Management's attitudes toward industrial supervision and their estimates of labor attitude. *Personnel Psychol.*, 1950, 3, 33-40.
5. Patton, W. M. A study of certain psychological variables related to supervision in the textile industry. Unpublished doctor's dissertation, Purdue University, 1951.



## An Application of Rogerian Concepts to Nurse-Patient Relationships

Lewis Bernstein<sup>1</sup>

*Veterans Administration Hospital, Denver, Colorado*

From the experience of teaching psychology to both student and graduate nurses, it has become increasingly apparent that psychologists can contribute in an important manner to nursing education. This potential contribution lies in the field of nurse-patient relationships, an area in their preparation which the nurses and student nurses themselves find incomplete. One nurse put the problem in this manner: "From the beginning of our training, the idea of caring for the patient, rather than the illness, has been emphasized, but nowhere in our program do we have the opportunity to learn how to put this idea into practice." Others have voiced a more specific need in such questions as: "I have a patient who has been in the hospital for two weeks and he has not yet had a visitor. He obviously feels uncomfortable and despondent during visiting hours. Is there anything I can do to make him feel better, or shall I ignore it?" "Patient X dies during the night and is removed from his room. On the following morning, other patients on the ward ask about X's whereabouts. Do we tell them the truth or use some subterfuge such as saying that he was transferred to another ward?"

These, and similar incidents, represent situations which nurses are called upon to handle, and for which they often feel unprepared. If nurses could understand the feelings expressed by their patients, and the motivation behind patients' behavior, not only would they feel more comfortable in these situations, but a real contribution to patient care and recovery would result.

In a study by Shields (6), schools of nursing, public health agencies, individual nurses, and other nursing groups were asked to in-

dicate, by means of a questionnaire, whether they thought that a basic nursing curriculum should provide learning experiences intended to develop certain qualities or abilities. One such quality was described as: "... a real belief in the essential worth of every human being and ... the importance of communicating this belief by attitudes and actions" (6, p. 12). This quality appears to be a direct translation of Rogers' concept of *reflection* which he defines as "... trying to understand from the client's point of view and to communicate that understanding" (5, p. 452). Although a large percentage of nurses who replied to the questionnaire felt that this was an important ability, some of the comments of the respondents reflect a doubt that such a quality can be taught. The following comments are among those reported by Shields: "A person either has or hasn't this quality. Shouldn't be a nurse if she hasn't it. Can't be taught (supervisor of a visiting nurse association)." "This comes with maturity and cannot be taught (private duty nurse)." "Criminals too? (private duty nurse)." "Idealistic. Impossible. No one can really believe in the essential worth of every human being (director of a school)." "Belongs in family teaching, not nursing education."

In view of such skepticism, it would seem worthwhile to determine empirically whether or not nurse-patient relationships using Rogers' concept of *reflection* can be successfully taught. This study, then, proposes to test the general hypothesis that nurses' skills and attitudes in interpersonal relationships can be modified in a significant fashion when the nurses understand the nature of the techniques they use, the attitudes which such techniques express or implement, and the feelings they generate in patients.

### Method

A series of three pretests, to be described below, was administered to all staff and head nurses on duty at the Denver VA Hospital.

<sup>1</sup>We wish to express our appreciation to Miss Marie L. Brophy, R.N., Chief Nurse, Miss Mary Jane A. McCarthy, R.N., Assistant Chief Nurse, and Miss Ruby L. Roepe, R.N., Assistant Chief, Nursing Education, all of the Veterans Administration Hospital, Denver, Colorado, without whose interest and cooperation this study could not have been completed.

Each nurse drew a number which was used as identifying information for the tests, thus providing personal anonymity. Upon completion of the pretests, it was announced that a hospital clinical psychologist was to conduct a course in nurse-patient relationships; that because of the size of the group, he would be able to work with only half of the nurses at one time. In order to obviate any feeling that some preference might be operating in the selection of those to participate in the course, the selection was made by the use of random numbers, in the presence of the entire group, using the same numbers with which they identified their pretests. This procedure provided an experimental group (those selected at random to participate in the course), and a control group (those not selected).<sup>2</sup> Table 1 indicates the degree of matching obtained by this randomization.

The course with the experimental group began approximately two weeks following the administration of the pretests. Ten weekly sessions of two hours each were held. The course began with a presentation of basic techniques which nurses use in responding to patients, a discussion of the attitudes which these techniques express, and a discussion of how the patient might react to each of these techniques. For the remainder of the course, nurses brought to class incidents from their own ward experience. These incidents were reported on a form which stated the situation and, as nearly verbatim as possible, the conversation that took place between the nurse and the patient. Behavioral responses accompanying the conversation were also reported. These incidents were discussed in terms of: (a) why did the patient behave in such and such a manner; what feelings was he really expressing by such verbalization and/or behavior; and (b) how could the nurse best respond in such a situation. Many of the incidents were role-played, and the implications of the situation discussed by the group. An effort was made to conduct the course along the lines suggested by Rogers' non-directive concepts (4, 5). That is, the instructor tried to create an atmosphere in which participants in the experiment could feel free to express all shades of opinion and criticism in the discussion of nurse-patient situations.

Upon completion of this training, both groups (experimental and control) retook the tests administered before the course was given.

### Measurement Techniques and Hypotheses

1. *The Nurse-Patient Situation Test.* This test is made up of 35 nurse-patient incidents,

<sup>2</sup> The tests were originally administered to 77 staff and head nurses. At the time of the posttesting, 59 of the original group were available—30 in the experimental group, and 29 in the control group. Eighteen subjects who took the pretests were either on leave or had resigned at the time of the posttests.

Table 1

The Degree of Matching Between the Experimental and Control Groups Achieved by Random Selection

Item	Experimental (N=30)	Control (N=29)
Median age	33.7	32.7
Mean no. years nursing experience	14.0	12.2
No. of graduates of hospital schools	28	26
No. of graduates of collegiate schools	2	3
No. who have received degrees since graduation from hospital school	7	7
No. of head nurses	5	6
No. of staff nurses	25	23
No. of medical nurses	10	10
No. of surgical nurses	9	11
No. of neurological nurses	5	4
No. of psychiatric nurses	3	2
No. of operating room nurses	2	1
No. of central supply room nurses	1	1

modified and adapted from Porter (3), together with five possible nurse responses to the statement of the patient. Each of the choices purports to measure one of the following five categories of response: E (Evaluative), H (Hostile),<sup>3</sup> S (Supportive), P (Probing), and U (Understanding).

A sample situation from the test, with the response choices, is the following:

I tell you I hate that doctor of mine. I hate him! I hate him! I ask him about my diagnosis and he gives me the brush-off. Tells me a diagnosis hasn't been made yet. Phooey! It makes me feel so terrible that I hate him so—especially when I have to count on him to get well. I—it worries me.

E (a) This is something you'll certainly want to get straightened out. A good relationship with your doctor is important for your recovery. You'll find he'll treat you better if you can make yourself have faith in him.

H (b) You're certainly not acting very grown-up. These doctors know their business. You do an awful lot of complaining about something that you're getting free.

S (c) I guess most patients go through a period when they don't like their doctors. It's really not at all uncommon. I hear that from most patients. But things eventually settle down.

<sup>3</sup> Porter (3) used the Interpretive category in place of our Hostile category. In an independent study of nurses' responses, we found the Hostile category used more frequently than the Interpretive, and that the few Interpretive statements made by nurses could be subsumed under the Hostile classification.



P (d) I think we ought to get at the root of that worry. Is there anything else your doctor has done to upset you besides not telling you your diagnosis?

U (e) You're concerned about how sick you really are, and it worries you not to know for sure what your doctor thinks.

The above example will also serve to illustrate the definitions of the five categories. In the Evaluative response, the nurse *has made a judgment of relative goodness of the patient's feelings*, and has *implied how the patient ought to feel and what he might do*. It would follow that the patient might not feel free to further explore his feelings about his physician since the nurse has, in effect, indicated his feelings are inappropriate.

The Hostile response in the above illustration again indicates to the patient the inappropriateness of his feelings and, in addition, *subjects him to ridicule* by implying that he is immature, and that he must accept whatever treatment is given him since the service is free.

Through the *reassurance* given the patient in the Supportive response, the nurse, in effect, *denies that the patient has a problem, that he need not feel as he does*. Although the denial of his feelings may preclude further discussion (leaving the nurse with the feeling that her reassurance has "worked"), it does not usually change the patient's feelings.

The Probing response implies that the patient *might profitably discuss the point further*, that if the patient will only *give her more information*, the nurse will be able to provide the answer or solution to his problem.

By means of the Understanding response, the nurse indicates that she is trying to *understand the patient's point of view, and to communicate that understanding to the patient*. The patient, feeling "safe" in such a situation, feeling that whatever attitudes he has are permissible, may now feel free to further explore, and himself modify, his feelings toward his physician. Furthermore, the patient, feeling that the nurse is an understanding person, may be generally more cooperative in other nursing procedures.

A try-out of this test on a class of 47 nursing students at the University of Colorado School of Nursing indicates that the five categories of response are relatively independent, with very little overlap. Intercorrelations between each category of response with every other category yielded low negative correlations with the exception of two non-significant low positive correlations. Furthermore, the test appears to be sufficiently reliable for use. Split-half reliabilities, correlating odd with even items, based upon the data of the 47 nursing students, are as follows for each category: Evaluative, .77; Hostile, .80; Supportive, .74; Probing, .88; and Understanding, .92.

*Hypothesis 1:* That the differences between

the posttest and pretest scores for the experimental group will show significantly greater decreases in all categories of response (except Understanding, which will show a significantly greater increase), than for the control group.

It is obvious that this Nurse-Patient Situation Test is a comparatively direct measure of the content of the course, but may not reflect a more basic change in underlying attitudes of the nurse. It was felt, therefore, that other independent measures of attitude change should be included in the test battery.

2. *The F-Scale.* As one independent measure of more basic changes in attitudes, the F-Scale was included in our test battery. This scale was developed in an extensive study of the "authoritarian personality," and is described in detail elsewhere (1). This scale measures attitudes on a continuum ranging from authoritarian to democratic.

*Hypothesis 2:* That the difference between the pretest and posttest scores for the experimental group will show a significantly greater shift toward the democratic end of the scale than for the control group.

3. *The Memory Test.* In this procedure, a lengthy case history, constructed from nurses' notes on an actual patient, is read to the group. The items in the history can be classified as physical items (temperature, blood pressure, diagnoses, laboratory procedures, medications, etc.), and psychological items (patient's ward behavior, the degree of his dependency, his employment history, etc.). Immediately after the history is read, the subjects are asked to write down everything they remember. The score on this test is:

$$\frac{\text{Number of physical items}}{\text{Number of psychological items}} \times 100.$$

A high ratio would indicate recall of more physical items than psychological. The rationale for using this procedure is that the case history is too long for the subjects to remember everything; that what is remembered will be selective; and that the course in nurse-patient relationships will make the experimental group more sensitive to psychological factors in the patient's history.

*Hypothesis 3:* That the difference between the pretest and posttest ratios between physical and psychological items will show a significantly greater drop for the experimental group than for the control group.

## Results

Table 2 presents the pre- and posttest scores for both groups, and the confidence levels of the pre-post differences between experimental and control groups.

The following facts are evident in Table 2:

1. The experimental group showed a significantly greater decrease in evaluative responses than the control group.



Table 2  
Differences Between Pretest and Posttest Scores

Test	Experimental		Control		P
	Pre	Post	Pre	Post	
Nurse-Patient Situation Test:					
Evaluative responses	10.0	1.0	12.5	11.0	.001
Hostile responses	1.0	.5	1.7	1.4	.90
Supportive responses	10.2	3.0	10.3	10.0	.001
Probing responses	9.7	2.2	8.2	9.5	.001
Understanding responses	4.5	28.3	2.3	3.6	.001
F-Scale	99.7	91.0	113.0	116.1	.01
Memory Test	173.4	144.1	173.0	193.0	.05

2. Neither group showed a significant decrease in hostile responses. The exceedingly small number of hostile responses by both groups (out of a possible total of 35) minimizes the importance of this category.

3. The experimental group showed a significantly greater decrease in supportive responses than the control group.

4. The experimental group showed a significant decrease in probing responses. The control group showed a slight increase in probing responses, although this increase is not significant.

5. The experimental group showed a significantly greater increase in understanding responses than the control group.

These data support our first hypothesis: that the experimental group would show a significantly greater decrease in evaluative, hostile (not significant), supportive, and probing responses, and a significantly greater increase in understanding responses than the control group.

6. The experimental group showed a significant shift in attitudes toward the democratic end of the F-Scale. The control group showed a slight, but not significant, shift toward the authoritarian end of the scale. These data support the prediction in hypothesis 2.

7. The experimental group showed a significantly lower ratio between physical and psychological items on the Memory Test. The control group showed a slightly higher, but not significant, ratio. These data support the prediction in hypothesis 3.

### Discussion

As previously stated, the attitudes and skills which we hoped to convey to the experimental group are based upon the nondirective concepts of Rogers (4, 5). It was, therefore, interesting to note that the group went through a process similar to that in a therapeutic counseling situation. Early in the course, many negative attitudes were freely expressed. As these were accepted by the instructor, more positive attitudes began to appear. The class itself noticed the conspicuous change in the "climate" of the course.

The question may arise if any more was accomplished than to train these nurses to recognize an understanding response. But the accompanying changes in sensitivity to psychological and social factors in a patient's case history, and the less authoritarian scores achieved on the F-Scale, do suggest that more basic changes took place. At a later date we plan to test the relative permanence of these changes. Several of the participants in the course were asked to explain their lowered scores on the authoritarian scale. The consistent response was that during the course they learned to respect the feelings of others, that patients could participate in the solution of their own problems, and that these attitudes could carry over to other spheres.

In addition to the changes in test findings there is other evidence that more than content was learned in the course. Nurse supervisors have reported that most nurses in the

experimental group are using these skills. Several of the group have requested additional training along the lines offered in the course. Even more convincing were the differences in understanding of patients' feelings noted between the incidents turned in for discussion early in the course and those submitted toward the end of the course.

This study has demonstrated that nurse-patient relationships making use of Rogerian concepts can be successfully taught. It is not to be inferred that a course such as that described in this paper is all that is necessary. Ideally, such a course should be taught early in nurses' professional education, and followed by appropriate ward supervision. The nurses represented in this study come from 59 nursing schools in 22 states. Yet, our pretest results indicate that very few had any meaningful preparation in this area. The study by Phillips and Agnew (2) indicates that the technique of giving understanding responses is "... considerably more than a simple extension of knowledge of interpersonal relations possessed by any reasonably intelligent and emotionally mature person." In other words, such skills and attitudes cannot be assumed to result from general nursing experience; they must be taught. And, with the current emphasis on interpersonal relations in nursing, the method herein described appears to be *one* manner in which such teaching may be accomplished.

#### Summary

Two groups of nurses—30 in an experimental group, and 29 in a control group—took a battery of three pretests. The Nurse-Patient Situation Test measured five categories of nurses' responses to patients' statements: Evaluative, Hostile, Supportive, Probing, and Understanding. The F-Scale measured social attitudes on a continuum ranging from authoritarian to democratic. The Memory Test measured the ratio of physical items to psychological items remembered from a lengthy case history.

Following the administration of the pretests, the experimental group participated in a course in nurse-patient relationships. An effort was made to conduct the course along the lines suggested by Rogers' nondirective

concepts. That is, the instructor tried to create an atmosphere in which participants could feel free to express all shades of opinion and criticism in the discussion of nurse-patient situations.

Upon completion of the course, both the experimental and control groups again took the series of tests described above, and the differences between the pre- and posttest scores were compared. On the Nurse-Patient Situation Test, the experimental group showed a significantly greater decrease in Evaluative, Supportive, and Probing responses, with a correspondingly greater increase in Understanding responses than the control group. No significant decrease in Hostile responses was demonstrated by either group. However, the exceedingly small number of Hostile responses minimizes the importance of this category.

The experimental group showed a significant shift toward the democratic end of the F-Scale, while the control group showed no significant change.

The ratio of physical to psychological items on the Memory Test showed a significant decrease for the experimental group, while the control group showed no significant change.

It is concluded that nurses' skills and attitudes in interpersonal relationships can be modified in a significant fashion when nurses understand the nature of the techniques they use, and the attitudes which such techniques express or implement, and the feelings they generate in patients.

*Received October 26, 1953.*

#### References

1. Adorno, T. W., Frenkel-Brunswick, Else, Levinson, D. J., and Sanford, R. N. *The authoritarian personality*. New York: Harper, 1950.
2. Phillips, E. L. and Agnew, J. W., Jr. A study of Rogers' 'reflection' hypothesis. *J. clin. Psychol.*, 1953, 9, 281-284.
3. Porter, E. H., Jr. *An introduction to therapeutic counseling*. New York: Houghton-Mifflin, 1950.
4. Rogers, C. R. *Counseling and psychotherapy*. New York: Houghton-Mifflin, 1942.
5. Rogers, C. R. *Client-centered therapy*. New York: Houghton-Mifflin, 1951.
6. Shields, Mary R. A project for curriculum improvement. *Nurs. Res.*, 1952, 1, 4-31.

## Instructor-Centered and Student-Centered Approaches in Teaching a Human Relations Course

Francis J. Di Vesta

*Syracuse University\**

The present study is a report on an experiment to evaluate: (a) the achievement of students in terms of outcomes desired from a human relations course; and (b) the relative effectiveness of two methods of teaching in achieving these outcomes. The course is one segment of a curriculum for the training of medical administrative supervisors in a military school. The experiment was conducted because of the lack of consistency in the findings from previous studies (1, 3, 11, 12, 13, 14, 17, 18, 19, 21, 22, 23). It is based in a general way on some of the procedures used in a previous study conducted by Canter (6). The present study differs from Canter's in that it compares two teaching methods, uses airmen rather than insurance company supervisors as subjects and incorporates a greater number of measures than used by Canter.

### Questions Studied

The present study was limited to a study of the following questions:

1. Is the twenty-hour block of instruction in human relations sufficient to increase the achievement level of students?
2. If changes are made by the instruction, what is the extent and direction of this change at the: (a) knowledge level; (b) attitudinal level; and (c) skill level?
3. Is one of the two methods of instruction more effective than the other for producing change in the student achievement level?

### General Procedure

The study was designed to take advantage of the best experimental procedure possible with a minimum disruption of course work and of routine generally employed in the conduct of the course. The over-all design was simply the pre-

test, instruction, post-test design and is shown below in more detail by steps.

*Step 1—Pre-Test:* All students in both the control and experimental groups were given all tests.

*Step 2—Instruction:* Students were divided into three groups for instruction purpose. Experimental group one received instruction by the instructor-centered method. Experimental group two received instruction by the student-centered discussion method. The third group was the control group and received only the technical instruction given in the course but did not receive the instruction given in human relations.

Students were selected for one or the other of the teaching methods on the basis of sociometric leadership ratings in a group performance test (see below for description of the test). Accordingly, those students from the first group to be administered the group performance test who were rated 1, 3, 5 were assigned to the instructor-centered method and those students who were rated 2, 4 and 6 were assigned to the student-centered method. Those students in the second group who were rated 1, 3 and 5 were assigned to the student-centered method and those who were rated 2, 4 and 6 were assigned to the instructor-centered method. This procedure was repeated until assignments had been made for all individuals. Those individuals who were assigned to the student-centered method were further sub-divided into sections of six people each for the actual instruction. These sub-groups were formed on the basis of a random sampling design which would assure that none of the individuals who were in the test groups worked together during the course. This procedure was a caution which assured the experimenter that one method would not have an advantage over another method on the group performance test as a result of an informal structure which might develop over a period of time during the formal course work.

*Step 3—Post-test:* All students in the experimental and control groups were given all tests. The testing situations and schedules were exactly the same for each individual on both the pre- and post-test.

During the course of the experiment two observers were placed with the class given instruction by the student-centered discussion method and one observer was placed with the class given instruction by the instructor-centered method. These observers checked on: (a) the extent to

\* This study was conducted while the author was on the staff of the Officer Education Division, Human Resources Research Institute, Maxwell Air Force Base, Alabama.



which instructional content material varied between the two classes and (b) the extent to which the instructor's approach was consistently oriented to the instructional method he was represented as using. Students in each class were also required to describe the instructional procedure through the use of a check list.

### The Criteria

The measures for criterion purposes were selected on the basis of the following requirements:

1. The test should measure some aspect of human relations ability or of leadership ability as established in previous studies.
2. The test should measure some aspect of school objectives.
3. The test should be dependable in its measurement properties.

4. The test battery should represent measures of human relations or leadership ability at the knowledge, attitude and behavioral levels (see below for fuller descriptions of these levels).

5. The tests should represent measures of objectives desired in the course.

The tests, classified according to level of measurement, are listed and described briefly below.<sup>1</sup>

**Knowledge tests.** What facts about human relations and leadership does the student know?

1. *Personnel relations test* (20). Developed by the Air Force's Human Resources Research Center for use with personnel in administrative positions. Measures what the student *knows* about supervisor-subordinate relationships.

2. *"How Supervise?" test* (7).

**Attitude tests.** How does he feel about certain kinds of leadership orientation?

1. *Problems of the Non-commissioned Officer in Charge* (4). A set of five scales developed and validated at Harvard University. Measures orientation toward discipline (severe-not severe); assessment of promotion practices (perceives much wrong-perceives little wrong); and handling of informal pressures in organization (try to satisfy these pressures-ignore pressures).

2. *Leadership Opinion Questionnaire* (8). Contains two scales. One measures orientation toward initiating structure in working with subordinates and aggressive directing of subordinates toward achieving the goal. The second scale measures the extent to which the supervisor is considerate of the feelings of those under him. Developed at Ohio State University.

**Skill tests.** The skill tests were used in an attempt to measure how the individual behaves in a realistic situation. These are divided into *Indirect* measures and *Direct* measures of behavior.

The indirect measures are paper and pencil

tests which establish situations for the respondent and require him to react to these situations. The advantage of this type of measurement is that it is possible to present a variety of situations to the respondent.

The direct measures are actual situations wherein the individual reacts to a realistic problem in conjunction with other individuals.

### Indirect Measures:

1. *Social intelligence test* (16).

2. *Prediction of human reactions test* (15). Developed for the Detroit Edison Company. It was revised for this experiment to be adaptable to the airman population. Primarily oriented toward judging how an individual would react given certain characteristics of individuals and circumstances which might occur in a supervisor-subordinate relationship.

### Direct Measures:

Students were assigned to a group composed of six airmen. They were told that they were to constitute a board to act on a morale problem occurring in a hospital staff. While acting as a board they were observed and rated by technicians trained for this kind of observation. Scoring was accomplished by using Bales's (2) interaction scoring form and by rating individuals according to the four roles of leadership activity, ability, likability and contribution of best ideas. Ratings on the four roles were also made by examinees at the close of the session as well as by observers. Sixteen groups in all were used.

### Instructors and Instructional Method

The instructors were selected because of their ability to use one of the methods. Each felt his competence was greatest in the method he was to use and was selected by his colleagues and superiors as being the most competent in that method of instruction. (Had it been possible, the experiment would have been replicated reversing the roles of the two instructors. However, before the next course began one of the instructors was separated from the service.)

Both instructors were carefully briefed on the content of the course. Each was warned not to emphasize content beyond that provided in the lesson plans. Observers were used to assure that the instructors remained within the content provided.

Decisions were made between both instructors and the experimenter as to how the instructional methods were to differ. Both observers and students rated the instructional methods on a check list of descriptive items. Items which, by item analysis using the chi-square statistic, were found to discriminate significantly ( $p < .05$ ) between the two methods of instruction are summarized qualitatively below. These items serve to describe the conduct of the teaching method as perceived by the students.

<sup>1</sup> Only the non-commercial tests are described. Detailed descriptions of the commercially available tests may be found in the references provided (5, 9).

Table 1

A Comparison Between Pre-Post-Test Scores for Control and Combined Experimental Groups

Criterion Measure		Control N = 24		Experimental N = 94	
		Mean	S.D.	Mean	S.D.
Personnel Relations Test	Pre	24.1	4.8	21.2	5.6
	Post	25.3	4.6	25.0	5.2
	t	1.84		9.43*	
<i>How Supervise?</i> Test Total	Pre	62.5	7.3	58.5	10.7
	Post	61.9	8.5	64.4	11.1
	t	.62		6.61**	
Supervisory Practice	Pre	12.3	3.2	11.8	2.8
	Post	11.7	3.2	12.3	3.3
	t	1.11		1.41	
Company Policies	Pre	24.4	3.9	23.5	4.9
	Post	25.4	3.5	24.2	4.7
	t	1.45		1.50	
Supervisory Opinion	Pre	25.8	3.9	23.2	5.8
	Post	24.8	4.6	27.9	6.1
	t	1.70		8.69**	
NCOIC					
Promotion-Orientation	Pre	2.4	1.2	2.8	1.0
	Post	2.2	.9	2.9	1.3
	t	.65		.42	
Assessment of Rewards	Pre	3.1	.9	3.5	1.2
	Post	2.4	1.2	3.1	1.2
	t	3.74**		3.08**	
Informal Pressures	Pre	3.0	1.4	2.6	1.1
	Post	2.7	1.2	2.8	1.1
	t	.94		1.53	
Discipline-Justice	Pre	2.8	1.2	2.4	.9
	Post	2.6	1.2	2.1	.9
	t	.67		2.58**	
Discipline-Initiative	Pre	3.4	1.1	3.5	1.1
	Post	3.2	1.4	2.9	1.3
	t	.84		4.21**	
Leadership Opinion Questionnaire Initiating Structure	Pre	52.0	7.7	54.8	7.3
	Post	48.5	7.5	52.0	6.6
	t	2.75*		4.38**	
Consideration	Pre	58.0	7.0	56.3	8.0
	Post	56.5	7.4	60.0	6.8
	t	1.52		5.83**	
Social Intelligence					
Judgment in Social Situations	Pre	20.7	3.2	19.6	3.8
	Post	21.2	3.5	20.0	2.3
	t	1.19		2.32*	
Observation of Human Behavior	Pre	37.0	7.8	33.0	11.7
	Post	40.7	8.1	36.6	10.2
	t	2.79*		6.14**	
Prediction of Human Reactions	Pre	48.6	10.3	47.7	11.2
	Post	47.2	11.1	50.0	11.7
	t	.30		2.97**	

\* = p .05 or &lt;.05.

\*\* = p .01 or &lt;.01.

## Method A

## Instructor-Centered

Suggestions were evaluated by instructor who advised or led class to correct conclusion.

Techniques and steps for activities were given by the instructor.

Instructor (rather than student) considers and handles individual problems and questions.

The instructor is the focus of attention. Student to student attention happens rarely or occasionally.

## Method B

## Student-Centered Discussion

Instructor encouraged suggestions and used this procedure to stimulate class to carry out class activities themselves.

Techniques and steps for activities emerged from the group discussion.

Group consideration of individual problems is encouraged by the instructor.

The instructor is the focus of attention whenever the discussion or activity needs guidance or information; otherwise students directed their attention to one another.

## Results

Only the results obtained from a study of the written tests and of the sociometric data are reported here.

The first hypothesis tested was that the course, regardless of method of instruction, produced an improvement in student achievement level. A comparison of pre and post scores for the control group with pre and post scores for the group receiving instruction was made for each of the tests. This comparison is shown in Table 1.

Significant ( $p < .01$ ) changes were made by the course segment in the knowledges related to human relations and leadership skills. These changes are reflected in the *Personnel Relations* and total *How Supervise?* test scores. The changes in the *How Supervise?* sections on *company policies* and *supervisory practices* were not significant although these tests also measured knowledge. An inspection of these tests indicates, however, that the content area is more appropriate to an industrial situation and would not be covered in a military course.

Important changes were also made in student attitudes. The most significant change in the attitude area is reflected in the pre and post test scores of students on the *How*

*Supervise?* section on *supervisory opinion* and on the *Leadership Opinion Questionnaire* section on consideration for others. Another change is noted in the *NCOIC Problems*. Student responses indicated a more lenient attitude toward problems of discipline and promotion after having attended the course than they did prior to attendance. It is interesting to note that *all* groups (including the control group) made significant changes on the *Leadership Opinion Questionnaire* section on initiating structure. This change was in the direction of a less favorable attitude toward active directing and structuring of situations in which leadership might be demonstrated. Undoubtedly some of the change occurred as a result of practice effect; however, the implication might be made that this change has occurred as a result of being in the school setting. An interesting hypothesis is stimulated here that the informal school setting, wherever it may be, may have detrimental effect on attitudes toward active initiation of structure. It is doubtful that such a change is more than a temporary one, although this hypothesis, too, should be a subject for further investigation.

In the area of indirect measurement of human relations skills the course, in general, effected significant changes. Students taking the course made significant gains on the *Social Intelligence* test section on judgment in social situations, and the *Prediction of Human Reactions* test. All groups (including the control group) made significant gains on the *Social Intelligence* test section on observation of human behavior.

The pre-post correlations for control and experimental groups on each of the measures are shown in Table 2. Although the reliability of the measures used here was available from previous studies using these instruments, it was desired to obtain some indication of reliability when used with our subjects. The pre-post correlations for the control group reflect a measure of the test-retest reliability. These correlations are shown in Table 2 with similar correlations for the experimental group. Five of the measures had pre-post correlations of .75 to .81; five measures, correlations of .62 to .68; three measures, cor-



Table 2

Pre-Post Correlations for Control and Experimental Groups on Each of the Tests

Test	Group	
	Control* N = 24	Experi- mental N = 94
Personnel Relations Test	.75	.74
How Supervise? Test		
Total	.79	.69
Supervisory Practice	.68	.42
Company Policies	.62	.55
Supervisory Opinion	.76	.62
NCOIC Problems		
Promotion-Orientation	.34	.44
Assessment of Rewards	.61	.45
Informal Pressures	.29	.21
Discipline-Justice	.49	.41
Discipline-Initiative	.51	.35
Leadership Opinion Questionnaire		
Initiating Structure	.67	.61
Consideration	.77	.66
Social Intelligence		
Judgment in Social Situations	.81	.90
Observation of Human Behavior	.66	.87
Prediction of Human Reactions	.53	.72

\* The control correlations amount to a measure of reliability.

relations of .49 to .53; and two measures had pre-post correlations of .29 and .34.

Canter's (6) research, in some respects, was similar to the present one. His study was conducted with supervisors of three large insurance companies. A control group was used but only the lecture discussion method was used in his study. The course was the same length (20 hours) as the one used in the present study and the content was similar. A comparison of the results of both studies, on tests appearing in both studies, is shown in Table 3.<sup>2</sup>

The insurance company supervisors achieved higher average scores on both the

pre-test and post-tests than did the airman population, on both tests. However, the airman population made a significant ( $p < .01 > .001$ ) change in scores, as a result of the course, on the *Prediction of Human Reactions* test whereas a significant change was not reported for the insurance company supervisors. Similarly, the airman population made gains: (a) *as great* as the insurance company supervisors on the "How Supervise?" *total score*; and (b) *greater* than the insurance company supervisors on the "How Supervise?" *supervisory opinions* score. On the other hand, Canter reports a significant change on the "How Supervise?" *company policy* score for the insurance company supervisors while the change for the airman population was not significant on this particular test.

*Knowledges, Attitudes and Indirect Measures of Skills.* As was noted earlier in this report, one of the weaknesses of this part of the study was that it was impossible to reverse the roles of the two instructors. However, it is assumed that each of the instructors was the best that could have been obtained for using the particular method. Students in each group rated their respective instructors the same way with regard to interest of the instructor in the subject. They described their respective instructor as "being interested in the academic progress of the students and interested in the students as individuals." The gains for the experimental groups, contrasted with the gains for the control group, on each of the tests, are shown in Table 4. The F test was used for testing significance of differences in gains for all groups. Where a significant F was found, the t test was applied between groups. Significant differences were found by the t test to occur only between the experimental and control groups.

In general, the evidence does not point to either method of instruction as being superior to the other. There was a general tendency, however, for the students taught by the lecture method to make greater gains on the knowledge and attitude tests than those students taught by the discussion method. These differences in gains made by the ex-

<sup>2</sup> To reduce printing costs Tables 3, 4, 5, and 6 have been deposited with the American Documentation Institute. Order Document 4323 from the ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress, Washington 25, D. C., remitting in advance \$1.25 for 35 mm. microfilm or \$1.25 for 6 x 8 in. photocopies. Make checks payable to Chief, Photoduplication Service, Library of Congress.

perimental groups were not significant statistically.

**Leadership Skills.** The measurement of leadership skills was conducted by placing the examinees in a simulated board meeting. The purpose of this meeting was to act on a morale problem which occurred in a hospital. During the meeting the examinees were rated by observers using Bales's Interaction Process analysis. Students were provided twenty minutes to act on the problem and five additional minutes for summarizing the discussion. After the meeting was over the respondents ranked one another from 1 through 6 on leadership, guidance, and best ideas. The same procedure was followed on the post-test. (See the section on procedures for a further description of how individuals were assigned to sections.)

The intercorrelations between the sociometric rankings are shown in Table 5.

The change in rankings are shown in Table 6. This table is based on the average score of the individual. Those individuals with an *average* leadership score of 0 to 1.99 were placed in category I (most leadership ability). Those with an average score of 2.00 to 3.99 were placed in category II and 4.00 to 6.00 were placed in category III (low leadership).

There was a significant change ( $p > .02 < .05$ ) via the chi-square test of significance in the pre-post lecture group on leadership. This difference is attributed largely to the increase in category I individuals and the decrease in category II individuals. Five people were rated as I (high leadership) before instruction and 13 after instruction.

A comparison of these two tables, however, shows some other trends which, although not significant, are worthy of consideration. For the discussion group individuals originally assigned to the middle category ( $N = 27$ ), there was very little movement out of this category. There was, however, a considerable movement of the discussion people originally assigned to category III. Approximately 50% of these individuals improved their leadership scores on the post-test. This movement does not occur for students in the lecture group.

## Summary

1. A 20-hour block of instruction in human relations, as taught in a course for air-men made a significant change in student performance. Students, taken as a body, showed significant gains in achievement (as measured by the pre-test results compared with the post-test results) on the following tests: (a) Personnel Relations Test; (b) How Supervise? (Total Score); (c) How Supervise? (Supervisory Opinions); (d) Leadership Opinions (Consideration Score); and (e) Social Intelligence (Judgment in Social Situations).

2. Furthermore, as measured by certain of these same tests the course was as effective as a similar course given to the supervisors of three large insurance companies. Students in the Medical Administrative Supervisor's course made gains *as great* as the insurance company supervisors on the "How Supervise?" test.

3. The use of the discussion method of teaching appeared to have a slight advantage over the instructor-centered approach in improving leadership ability. There was a strong tendency for students starting the course at a low leadership level to improve through the discussion method. This tendency did not exist for individuals taught by the instructor-centered approach. Students at the upper levels of leadership ability are not affected much by either method. This finding was necessarily based on a small number of people. It should be made clear that a *tendency*, not a *clear-cut* change was found. A replication of the experiment would provide more definitive data.

4. As measured by the knowledge and attitude tests, there does not appear to be an advantage in the discussion approach over the instructor-centered approach. Both methods produced equally good results. It should be emphasized that this finding applies only to one way of using the discussion method. Other variations of the discussion method (or of the instructor method) could produce quite different results.

5. There is a pronounced and significant change in student attitude in general, toward

initiating structure. Students, after being in school, tend to feel that initiating structure in group situations is less important than they did before the course started. This change appears to occur by virtue of being in the *school situation* and is *not* directly attributable to a particular teaching method. The evidence that this occurs as a result of being in a school situation is that this change occurred for each group including the control group. Further research would be necessary if it is desirable to know whether this is a temporary change or a permanent one. It is doubtful that the change is permanent. Further research would also be required to yield answers about how to develop a school atmosphere that would promote positive attitude toward initiating structure.

Received October 29, 1953.

#### References

1. Asch, M. F. *Nondirective teaching in psychology: an experimental study*. *Psychol. Monogr.*, 1951, 65, No. 4 (Whole No. 321).
2. Bales, R. F. *Interaction scoring form*. Cambridge, Mass.: Addison-Wesley Press, Inc., 1949.
3. Bane, C. L. The lecture versus the class discussion method of college teaching. *Sch. & Soc.*, 1925, 21, 300-302.
4. Borgatta, E. *Questionnaire on problems of the non-commissioned officer*. Cambridge, Mass.: Harvard University, 1952.
5. Buros, O. K. *The fourth mental measurements yearbook*. New Jersey: Gryphon Press, 1953.
6. Canter, R. R. A human relations training program. *J. appl. Psychol.*, 1951, 35, 38-45.
7. File, Q. W. and Remmers, H. H. (Ed.) *"How Supervise?" Form A*. New York: The Psychological Corporation, 1948.
8. Fleishman, E. A. *Foreman's leadership opinion questionnaire*. Columbus, Ohio: Ohio State University, 1952.
9. Fleishman, E. A. *"Leadership climate" and supervisory behavior*. Ohio: Personnel Research Board, The Ohio State University, 1951.
10. Guetzkow, H. *Groups, leadership and men*. Pittsburgh: Carnegie Press, 1951.
11. Husband, R. A statistical comparison of the efficacy of large lecture versus smaller recitation sections upon the achievement in general psychology. *Amer. Psychologist*, 1949, 4, 216. (Abstract.)
12. Jones, H. E. Experimental studies of college teaching. *Arch. Psychol.*, 1923, 10, No. 68.
13. Katzell, R. A. Testing a training program in human relations. *Personnel*, 1946, 23, 85-97.
14. Krech, D. and Crutchfield, R. S. *Theory and problems of social psychology*. New York: McGraw Hill Book Co., Inc., 1948.
15. Meyers, H. H. *Human relations—A test of ability to predict human reactions*. New York: H. H. Meyers, 1952 (Revised).
16. Moss, F. A., Hunt, T., and Omwake, K. T. *Social intelligence test, SP edition*. Washington, D. C.: George Washington University, 1947.
17. Roseborough, Mary E. Experimental studies of small groups. *Psychol. Bull.*, 1953, 50, 275-303.
18. Sanford, F. H. and Hemphill, J. F. *An evaluation of the text "Psychology for Naval Leaders used in leadership training at the naval academy"*. College Park: University of Maryland, Department of Psychology, 1948.
19. Spence, R. B. Lecture and class discussion in teaching educational psychology. *J. educ. Psychol.*, 1928, 19, 454-462.
20. Zacarria, M. A. *Personnel relations test*. San Antonio, Texas: Human Resources Research Center, 1952.
21. *Development of evaluative and predictive measures in the Air Force Officer Candidate School*. Washington, D. C.: Human Resources Research Commission Research Bulletin 47-1, 1947.
22. *Research on military leadership*. Washington, D. C.: Panel on Human Relations and Morale, Committee on Human Resources, Research and Development Board, 1951.



## Vocational Interests and Socio-Economic Status

John W. Gustad

*University of Maryland*

Prominent among the factors which are thought to influence the choice of an occupation is socio-economic status. This may include both the status accorded to the occupation by others as well as the level of aspiration of the individual concerned. It is this latter, the level of aspiration of the individual with respect to occupations, that is the concern of the present investigation.

Early in his work, Strong (9) recognized the need for a measure of the status aspirations of individuals completing his interest blank. He accordingly developed a scale which he called Occupational Level (henceforth referred to as OL). This was accomplished by contrasting the item responses of laboring men with those of men in business or the professions earning over \$2500 per year. It should be noted that at the time this scale was built, this figure represented the upper fifth of the income distribution in this country.

Since its publication, a considerable amount of research has been done on and with OL. It has seemed promising as a measure of motivation, level of aspiration, or socio-economic status drive. Darley (2, p. 60) has called it "... a quantitative statement of the eventual adult level of aspiration." Darley (2), Gustad (5), Kendall (6), Ostrom (7, 8), and Strong (9) have shown that OL has some relationship to success or staying power in college.

In an extensive study of OL as a measure of drive, Barnett *et al.* (1) reported several interesting relationships. OL was found to correlate .44 with a self-rating of level of aspiration in one school, .04 in another; with a verbal level of aspiration measure, .26 in the first school, .18 in the second. Though the results were not entirely clear, it was concluded that there was some relationship between OL and other measures of level of aspiration. Stewart, in the same monograph, concluded that "... the mother may have a greater influence on the development of voca-

tional interests than has hitherto been assumed" (p. 17). It should be noted, however, that the sample studied was composed of the sons of skilled workmen.

Recently, Gough (3, 4) has developed two scales for measuring different aspects of socio-economic status. One is essentially a shortened, more easily administered version of scales used to assess actual, objective status. The other attempts to get at the individual's level of aspiration regardless of his objective status. These will henceforth be referred to as objective and subjective status respectively.

### The Problem

The present study was designed to answer two principal questions: first, how, if at all, do various interest groups differ in terms of socio-economic status, however defined; second, what are the relations among the various measures of status, all of which were designed to get at a common variable in different ways?

The subjects were all men students in the junior classes of the colleges of Arts and Sciences and Engineering at Vanderbilt University. Men were selected both because of the generally better understanding of their interests as well as for the fact that there is no OL key on the women's form of the Strong.

All subjects completed the Strong Vocational Interest Blank as well as the two scales devised by Gough. Interest blanks were scored for all thirty-nine occupational keys as well as for the three clinical keys, OL, Interest Maturity, and Masculinity-Femininity. Interest profiles were sorted in accordance with the method outlined by Darley (2) into primary interest groups. Those subjects who had no primary patterns were retained for study as a separate group (N.P.). Twenty-six cases, approximately ten per cent of the sample, had more than one primary. Examination of the profiles showed that in all

but four cases one primary might be considered to be stronger or "more primary" than the other and was accordingly chosen. In the remaining cases, secondary and tertiary patterns were inspected and a judgment made in favor of one or the other primary in terms of the total configuration. Those areas in which the subjects had primary patterns were as follows: Biological Sciences, Physical Sciences, Sub-technical, Social Welfare, Business Detail, Sales, and Verbal-Linguistic.

After  $L_1$  tests indicated homogeneous variances, analyses of variance were made for each status measure across all interest groups. Product-moment correlations among the status measures were also computed.

### Results

The results of the analyses of variance are shown in Table 1. Of the three status measures, only OL showed significant differences among interest groups.

To investigate further the situation with regard to group differences on OL, tests of the significance of the differences between all means were made. These are included in Table 2. While there were scattered significant differences in several groups, the two groups which appeared to be most consistently different were the Sub-technical and Verbal-Linguistic. The OL scores of the former tended to be below average for the present sample while those of the latter were above average. These results are in close agreement with those reported by Strong (9)

who correlated OL scores with scores on individual scales.

Finally, the correlations among the three measures were computed. They were as follows: OL and subjective status, .07; OL and objective status, .10; objective and subjective status, — .03. None of these was statistically significant. Gough (3) reported a correlation of .52 between his two scales in a sample of high school seniors.

### Discussion

From the foregoing, it must be concluded that at least for the present sample there is independence among the three status measures and only significant differentiation among interest groups in the case of OL. Several possibilities may account for these findings.

In the first place, only OL was specifically built to differentiate among occupational groups, but even it was not directly related to specific interest groups or occupations. Yet Barnett *et al.* (1, p. 13) say that "The OL scores may be hypothesized as reflecting the individual's socio-economic goals in life." Further, on p. 17, they say, "The OL score is so constructed that it should indicate the socio-economic level of an individual's interests." In many ways, the development of OL was quite similar to that of subjective status; both involve self-descriptions about preferences for activities, reactions, feelings, etc.

Another possibility lies in the nature of the sample which in the present case was drawn from students attending a private, fairly expensive, above average socio-economic status university. These men were for the most part preparing for jobs in the professions or in business management. This is in direct contrast to the sample used by Stewart (1) described above. There may have been a ceiling effect operating to restrict the range near the upper limit. Yet if this were the case, such an effect should presumably have operated on the other scales in the same way as on OL which did not happen.

It may be that OL is a more specific-to-occupations kind of measure than the other two scales. This should be studied, but the manner of development of all three makes

Table 1  
Analyses of Variance of Status Measures  
Across Interest Groups

Variance Source	Status Measure		
	OL	Objective Status	Subjective Status
Between	397.96	22.75	22.83
Within	18.66	11.48	8.17
Total	416.62	34.23	31.00
F*	21.33	1.98	2.79
P	<.001	>.05	>.05

\* Degrees of freedom for all three measures were as follows: for Between, 7; for Within, 244; Total, 251.

Table 2  
Mean Differences on Occupational Level for All Interest Groups

Interest Group	Mean	N	Interest Group						
			Nat. Sci.	Sub.-Tech.	Soc. Welf.	Bus. Det.	Sales	Verb. Ling.	N.P.
Bio. Sci.	54.73	26	1.53	7.04**	2.88**	1.43	-1.58	- 5.83**	.18
Nat. Sci.	53.20	24		5.51**	1.35	- .10	-3.11*	- 7.36	-1.35
Sub. Tech.	47.69	54			-4.16**	-5.61**	-8.62**	-12.87	-6.86**
Soc. Welf.	51.85	20				-1.45	-4.46**	- 8.71**	-2.70*
Bus. Det.	53.30	23					-3.01*	- 7.26**	-1.25
Sales	56.31	49						- 4.25*	1.76
Verb.-Ling.	60.56	9							6.01**
N.P.	54.55	47							
		252							

\* Denotes significant at or beyond the .05 level.  
\*\* Denotes significant at or beyond the .01 level.

this appear unlikely. The study cited above (1) is again pertinent. Gough's objective status scale probably gives greatest weight to factors contributed by the father. If Stewart's results may be accepted, the interest group in which the individual is finally found is more a function of maternal influence.

What is probably needed is more work on the nature and dimensions of vocational interests as well as on socio-economic status. There are some contingencies, for instance, which should be considered. An individual from a high status home might have what is for him a low status score and yet still be average or above. Similarly, another person from a low status home might have what for him is a very high status score and he too might be average.

Conclusions

1. Of the three status measures studied, only OL differentiated significantly among the interest groups.
2. Study of the mean differences with respect to OL showed that those individuals with Sub-technical interests tended to have below average OL scores while those with Verbal-Linguistic interests tended to be above average on OL.

3. There was no significant correlation among the three status measures in the present sample.

Received September 25, 1953.

References

1. Barnett, G. J., Handelsman, I., Stewart, L. H., and Super, D. E. The Occupational Level scale as a measure of drive. *Psychol. Monogr.*, 1952, 66, No. 10 (Whole No. 342).
2. Darley, J. G. *Clinical aspects and interpretation of the Strong Vocational Interest Blank*. New York: Psychological Corporation, 1941.
3. Gough, H. G. A short social status inventory. *J. educ. Psychol.*, 1949, 40, 52-56.
4. Gough, H. G. A new dimension of status: I. Development of a personality scale. *Amer. sociol. Rev.*, 1948, 13, 401-409.
5. Gustad, J. W. Academic achievement and Strong Occupational Level scores. *J. appl. Psychol.*, 1952, 36, 75-78.
6. Kendall, W. E. The Occupational Level key of the Strong Vocational Interest Blank for Men. *J. appl. Psychol.*, 1947, 31, 283-287.
7. Ostrom, S. R. The OL key of the Strong Vocational Interest Blank for Men and scholastic success at the college freshman level. *J. appl. Psychol.*, 1949, 33, 51-54.
8. Ostrom, S. R. The OL key of the Strong test and drive at the twelfth grade level. *J. appl. Psychol.*, 1949, 33, 241-248.
9. Strong, E. K., Jr. *The vocational interests of men and women*. Stanford: Stanford University Press, 1943.



## Permanence of Interests and Interest Maturity<sup>1</sup>

Kalmer E. Stordahl

*Arkansas Polytechnic College, Russellville, Arkansas*

Many counselors in working with college and precollege youth use Strong's Vocational Interest Blank. In using this blank, or any other interest inventory, they are concerned with the problem of the permanence of scores obtained on the blank. The Strong blank has a scale, Interest Maturity, which is used by many counselors as a measure of the probable stability of a counselee's interest profile. They assume a positive relationship between stability of interests and Interest Maturity score. There is, however, very little or no evidence to support this assumption. For an account of how the Interest Maturity scale was constructed and the evidence for its relationship to permanence of interests, see Strong (4). The present study was designed to test whether or not scores on the Interest Maturity scale are related to interest stability.

### Method

In 1949 the Vocational Interest Blank was offered on an optional basis to all high school seniors who participated in the state-wide testing program in the State of Minnesota. Approximately 3500 senior boys completed the blank. These completed blanks were made available to the investigator by the Student Counseling Bureau at the University of Minnesota.

A check was made of the University of Minnesota enrollment in 1951 to determine how many of these boys were enrolled at that time. It was found that 331 boys who had completed the Strong in 1949 were enrolled. A sample of 206 of these boys was contacted and asked to again complete the Strong blank; 182, 88 per cent, complied with this request. One subject omitted a number of items making his blank unusable so that tests and retests for 181 subjects were used in the study.

The minimum time between test and retest was two years and the maximum time did not exceed 2.5 years. The mean age of the 181 subjects at the time of the retest was 19.8 years.

The tests and retests for the 181 subjects were

scored for forty-four occupational scales and for Interest Maturity, Occupational Level and Masculinity-Femininity. To determine whether Interest Maturity score was related to stability of interests, some measure of stability for the individual was needed. Kendall's (2) coefficient of concordance,  $W$ , was used for this purpose. The coefficient  $W$  is based on the method of ranks. It is related to Spearman's  $\rho$  but has the advantage of being appropriate for any number of observations, whereas  $\rho$  is applicable only to two sets of data. In the present study,  $\rho$  would also have been appropriate as only two sets of data were used.

Coefficients of concordance were computed between each individual's test and retest profile. The forty-four occupational scales were used in computing this coefficient.

The subjects' interest profiles were arbitrarily divided into three groups of approximately equal size on the basis of their  $W$  values. Those with coefficients of concordance between test and retest of .906 to .977 were designated as a "high" stability group ( $N=60$ ), those with concordance values of .820 to .905 were designated as an "average" stability group ( $N=61$ ), and those with concordance values of .419 to .818 were designated as a "low" stability group ( $N=60$ ). The Interest Maturity scores of these three groups on the first test, i.e., the 1949 test, were then compared.

### Results

The coefficients of concordance are of interest themselves as a measure of the stability of individual Strong profiles. The range of coefficients was from .42 to .98 with a median of .87. All but fifteen of the 181 coefficients were significantly greater than zero at the .01 level. Since  $W$  has a direct relationship to Spearman's  $\rho$ , these figures can also be expressed in terms of  $\rho$ . The median  $\rho$  would be .74.

The means and standard deviations of the Interest Maturity scores for the "high," "average," and "low" interest stability groups are given in Table 1. Bartlett's test for homogeneity of variance indicated that the variances were homogeneous ( $P > .05$ ). An analysis of variance of the Interest Maturity scores (Table 2) showed that no significant

<sup>1</sup> This paper is based upon a portion of a Ph.D. thesis submitted to the graduate faculty of the University of Minnesota. The author wishes to acknowledge the guidance of his advisor, Dr. Willis E. Dugan.

Table 1

Means and Standard Deviations of the Interest Maturity Scores for Subjects with High, Average, and Low Coefficients of Concordance

W value	N	Mean	S.D.
High (.906-.977)	60	46.7	8.8
Average (.820-.905)	61	48.6	7.0
Low (.419-.818)	60	46.5	8.9

difference existed between the means of the three stability groups ( $P > .05$ ).

Although the Interest Maturity scores on the first test did not differ for the three stability groups, the mean Interest Maturity score for the entire sample of 181 increased from 47.2 on the first test to 52.0 on the retest. This increase was significant at the .01 level.

These results fail to substantiate the assumption of a positive relationship between interest stability and Interest Maturity score. From this, one may conclude that the present Interest Maturity scale is not useful as a means of estimating the probable stability of a precollege male's interest profile. More useful would be a key built by contrasting the responses of persons whose interests remain stable with the responses of persons whose interests do not remain stable over a period of time.<sup>2</sup>

### Summary

A sample of 181 males who had completed Strong's Vocational Interest Blank as high school seniors were retested two years later as college students. Using Kendall's coef-

<sup>2</sup> The Student Counseling Bureau at the University of Minnesota has begun work on such a key.

Table 2

Analysis of Variance of Interest Maturity Scores for Subjects with High, Average, and Low Coefficients of Concordance

Source	df	SS	MS	F	P
Between	2	159.73	79.865	1.177	>.05
Within	179	12142.09	67.833		
Total	181	12301.82			

Note: Bartlett's test for homogeneity of variance: chi square = 4.17;  $P > .05$ .

ficient of concordance, W, as a measure of the relationship between the test-retest profiles, coefficients were computed for each of the 181 pairs of profiles. When those individuals with high ( $N = 60$ ), average ( $N = 61$ ), and low ( $N = 60$ ) W values were compared with respect to Interest Maturity score on the first test, they were found to be homogeneous. Thus, the results of this study do not support the assumption of a positive relationship between interest stability and Interest Maturity score.

Received October 1, 1953.

### References

1. Johnson, P. O. *Statistical methods in research*. New York: Prentice-Hall, 1949.
2. Kendall, M. G. *The advanced theory of statistics*, Vol. I. (4th Ed.) London: Charles Griffin & Co., 1948.
3. Stordahl, K. E. The stability of Strong Vocational Interest Blank patterns for pre-college males. Unpublished doctor's dissertation, University of Minnesota Library, 1953.
4. Strong, E. K., Jr. *Vocational interests of men and women*. Stanford: Stanford Univer. Press, 1943.

# The Strong Vocational Interest Blank and College Achievement

Ralph M. Rust and F. J. Ryan

*Division of Psychiatry and Mental Hygiene, Department of University Health,  
Yale University*

An awareness that prediction of college grades by means of "intellectual" factors had reached a point of diminishing returns stimulated extensive efforts to explore the predictive value of other personality variables. A number of authors (2, 3, 11) have reviewed these attempts. Thus far, the coefficient of alienation left by the present predictors has been little reduced.

The Strong Vocational Interest Blank for Men (SVIB), because of the reliability of its scales and its wide usage, has been included in much of the research on academic achievement. Most of these studies have been summarized by Strong (10). A diversity of designs, definitions of achievement, and instruments, often along with methodological deficiencies, renders interpretation of results difficult. It does appear, however, that keys for the Strong (e.g., 12, 13) have been developed which can add to the prediction of college grades afforded by intelligence test scales. Yet, such scales fell into disuse through their failure to add to a predictive battery which includes secondary school grades.

In the authors' (8, 9) preliminary investigation of personality variables associated with academic achievement, the different definition of achievement used appeared to warrant a re-examination of this factor's relationship to SVIB.

At present, the best available estimate of an incoming freshman's grades at Yale College is his "general predicted score" (1). This measure is the dependent variable in a multiple regression equation of which the three independent variables are: (1) adjusted secondary school record; (2) Scholastic Aptitude Test score (SAT); and (3) the total of three College Entrance Board Examinations (CEEB). Achievement was measured in terms of deviation from predicted score. Thus, unlike most previous studies, the investigation was concerned with achievement beyond that predicted by a battery which includes secondary school record. A recently reported study by Melville and Frederiksen (5)

on freshman engineering students at Princeton also used adjusted secondary school grades as one component of the predicted score.

Though the results yielded by the first experimental groups (Yale College classes of 1950 and 1951) gave little promise that the Strong would have practical prediction value for academic achievement, the scales showed more than a chance relationship to achievement status. Further, the significant results obtained for Group V scales and the Masculinity-Femininity (M-F) scale appeared to offer indirect support for a hypothesis developed earlier (8, 9). For these reasons, SVIB was included in a battery administered to three other sets of experimental groups. Data were available for an additional group. These extensive replications permit an improved appraisal of SVIB as it relates to college academic achievement.

## Subjects

Selection of subjects was based on the relationship between grades and predicted scores. The procedure, described elsewhere in greater detail (8, 9) was designed to yield three groups of Yale undergraduates who would be equated for predicted score, but who would differ widely in grades. The regression line of grades on predictions was drawn on a scattergram. Lines parallel to the regression line were drawn so as to cut off approximately the most extreme ten per cent of both the positive deviants from predicted score, over-achievers (O's), and the negative deviants, underachievers (U's). A third group, normal-achievers (N's), included those students in cells cut by the regression line. This procedure was applied to each sample separately. Subjects included in four samples had accepted invitations to participate in the study and were tested in either the junior or senior year. In one sample, students were routinely administered the test during the freshman year.<sup>1</sup>

Table 1 indicates that the combined experimental groups do not differ in predicted scores

<sup>1</sup> Strong scores for this group were obtained by J. R. Wittenborn.



Table 1

Comparison of Groups on Prediction, Components of Prediction, and Average Grade

	Prediction		School Grade Adjusted		SAT		Average CEEB		Average Grade		N
	M	SD	M	SD	M	SD	M	SD	M	SD	
U	76.9	4.9	76.5	5.9	58.7	8.3	57.8	5.6	71.4	3.5	143
N	77.1	4.9	76.5	5.8	59.0	8.2	57.5	6.5	77.9	3.2	175
O	77.4	4.9	77.2	5.6	59.8	8.7	57.7	6.1	84.3	3.4	165
T	77.1	5.0	76.7	5.8	59.2	8.4	57.6	6.1	78.2	6.1	483
CR <sub>UN</sub>	.3		.1		.3		.5		17.1		
CR <sub>UO</sub>	.8		1.0		1.2		.2		32.2		
CR <sub>NO</sub>	.6		1.2		.9		.3		18.3		

or components of predicted scores, but differ significantly in academic average.

### Results

The ability of SVIB occupational scales to separate the experimental groups is shown in Table 2. Comparisons among groups were made by means of chi-square with the cut-off point for each scale taken at the median of the total group. None of the 44 scales differentiated U's from N's. Overachievers differ significantly ( $p < .05$ ) from U's on 11 scales, and from N's on 12 scales, nine of these scales being the same. These significant differences are as follows: overachievers score higher than both other groups on scales for Artist, Psychologist, City School Superintendent, Minister, Musician and C.P.A. They also score higher than N's on Mathematician, Group II, and Group X. Overachievers score lowest on Sales Manager, Real Estate Salesman, and M-F. In addition, they score lower than U's on Aviator and Forest Service.

A key<sup>2</sup> for each achievement group was developed from item analyses of two samples. Uncorrected odd-even reliabilities are .43 for the U key, .38 for the N key, and .42 for the O key. Results of the application of these keys to the remaining three samples combined are shown in Table 3. Results of subtracting each subject's U score from O score (O-U

score) are also shown. The U key and the O-U score yield significant differences between O's and U's, and O's and N's, whereas the O key differentiates only O's from U's. The N key produces no significant differences. None of the keys yields significant differences between U's and N's.

A measure of the congruency between stated occupation choice and Strong scores was available for 265 subjects. The percentage of each group receiving A scores on the Strong in their occupational choice is: U's, 36.7; N's, 34.3; and O's, 32.2. The differences among groups are not significant.

### Discussion

Though a clear interpretation of results is hampered by the empirical nature of the instrument, two aspects compel attention. The occupational scales apparently distinguish among the achievement groups with more than chance frequency and consistency.<sup>3</sup> Scoring keys, empirically developed from two samples, separated the remaining samples with statistical significance. Both of these events can be viewed as evidence that achievement as measured in this study is not an artifact produced by the unreliabilities of either the predictors or grades. Further, evidence is supplied that there is a relationship between achievement status and responses to the Strong items.

The ability of empirical scoring keys of low reliability to separate the experimental groups offers some promise for eventual development of keys which would add signifi-

<sup>3</sup> Since scale scores cannot be treated as independent events, estimates of chance expectancy can only be approximately determined.

<sup>2</sup> To reduce printing costs, these keys and the method used to select items for the keys, have been deposited with the American Documentation Institute. Order Document No. 4324 from the ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress, Washington 25, D. C., remitting in advance \$1.25 for 35 mm. microfilm or \$1.25 for 6 x 8 in. photocopies. Make checks payable to Chief, Photoduplication Service, Library of Congress.

Table 2  
Achievement Status and Strong Vocational Interest Blank Scores

Scale	Per Cent above Median			Chi-Squares Significant at .05 Level		
	Under- achievers N = 139	Normal- achievers N = 175	Over- achievers N = 166	U vs. N	U vs. O	N vs. O
Artist	46.0	43.4	60.8	— (3-2)	6.67(5-0) <sup>1</sup>	10.35(4*-1)
Psychologist	43.9	40.6	61.4	— (3-2)	9.38(5*-0)	14.85(5*-0)
Architect	53.2	46.3	53.6	— (3*-2)	— (3-2)	— (4-1)
Physician	49.6	46.3	53.6	— (4*-1*)	— (3-2)	— (4*-1)
Dentist	43.8	46.9	53.6	— (2-3)	— (4-1)	— (4-1)
Group I <sup>2</sup>	46.5	48.7	54.1	— (3-1)	— (4-0)	— (3-1)
Mathematician	48.9	42.3	57.0	— (4-1)	— (4-1)	7.33(5*-0)
Engineer	54.7	48.6	44.0	— (4*-1)	— (5-0)	— (3-2)
Chemist	50.4	48.0	50.0	— (4-1)	— (3-2)	— (4-1)
Group II <sup>2</sup>	46.5	41.9	57.9	— (3*-1)	— (3*-1)	6.39(3*-1)
Prod. Mgr.	55.4	51.4	44.6	— (3-2)	— (4-1)	— (3*-2)
Aviator <sup>2</sup>	55.3	52.1	42.1	— (3-1)	4.26(4-0)	— (3-1)
Farmer <sup>2</sup>	53.5	53.0	42.1	— (1-3)	— (3-1)	— (3-1)
Carpenter <sup>2</sup>	50.9	50.9	40.6	— (2-2)	— (3*-1)	— (3*-1)
Printer <sup>2</sup>	51.8	49.6	49.6	— (2-2)	— (2-2)	— (2-2)
Math.-Sci. Teacher	55.4	47.4	48.8	— (4-1)	— (4*-1)	— (3-2)
Policeman <sup>2</sup>	51.8	52.1	46.6	— (2-2)	— (3*-1)	— (3-1)
Forest Service <sup>2</sup>	56.1	50.4	43.6	— (4-0)	3.84(4*-0)	— (2*-2)
YMCA Phys. Dir.	46.0	46.3	53.6	— (3-2)	— (4-1)	— (3*-2)
Personnel Dir.	47.5	49.7	51.8	— (3-2)	— (3-2)	— (1*-4)
YMCA Secretary	45.3	47.4	53.9	— (3-2)	— (4-1)	— (2-3)
Social-Sci. Teacher	43.8	48.0	54.8	— (3*-2)	— (5*-0)	— (2*-3)
City School Supt.	41.7	46.3	61.4	— (4*-1)	11.80(5**-0)	7.87(5*-0)
Minister	45.3	45.7	60.8	— (2-3)	6.67(5-0)	7.83(4*-1)
Group V	43.2	50.3	54.2	— (4*-1)	— (4-1)	— (2*-3)
Musician	46.0	46.9	57.8	— (2-3)	4.22(5-0)	4.11(4-1)
C.P.A.	46.0	43.4	59.4	— (2-3)	5.42(3.5*-1.5)	8.66(3*-2)
Accountant	46.0	53.1	52.4	— (3-2)	— (4-1)	— (2-3)
Office Worker	50.4	52.0	47.0	— (4-1*)	— (2.5-2.5)	— (4-1)
Purchasing Agent	49.6	52.6	46.4	— (3-2)	— (4-1)	— (4-1)
Banker <sup>2</sup>	47.5	56.4	47.4	— (3-1)	— (1.5-2.5)	— (3-1)
Group VIII <sup>2</sup>	51.8	52.1	48.1	— (3-1*)	— (2*-2)	— (3-1)
Sales Manager	55.4	55.4	44.0	— (3-2) <sup>3</sup>	3.95(5*-0)	4.47(4-1)
Real Estate Salesman	54.7	57.7	42.8	— (3-2)	4.29(5-0)	7.61(5**-0)
Life Insur. Salesman	50.4	49.1	45.8	— (2-3)	— (4-1)	— (3-2)
Group IX <sup>2</sup>	49.1	53.0	41.4	— (3-1)	— (4-0)	— (3-1)
Advertising Man	48.2	47.4	52.4	— (3-2)	— (3*-2)	— (2*-3)
Lawyer	46.8	48.6	57.8	— (3*-2)	— (3*-2)	— (4*-1)
Author-Journalist	48.2	47.4	57.8	— (4-1*)	— (3*-2)	— (3*-2)
Group X <sup>2</sup>	43.0	41.0	54.1	— (3-1)	— (2**-2)	4.29(2**-2)
President <sup>2</sup>	49.1	52.1	46.6	— (2-2)	— (3-1)	— (3-1)
Occupational Level <sup>2</sup>	48.2	56.0	59.1	— (3-1)	— (3*-1)	— (2*-2)
Masc.-Fem.	56.8	57.7	40.1	— (3-2*)	8.23(4*-1)	10.26(5*-0)
Interest Maturity <sup>2</sup>	43.4	48.7	55.6	— (3-1*)	— (4-0)	— (1*-3)

<sup>1</sup> The first number in parentheses gives the number of times the direction of the results of the samples was the same as that of the combined group. The second number indicates reversals. Asterisks are added for each sample significant at the .05 level or .01 level or better.

<sup>2</sup> Comparisons based on four samples. Number of subjects: U = 114; N = 117; O = 133.

<sup>3</sup> Normalachievers exceed underachievers in three samples.

Table 3  
Application of Achievement Keys to a New Sample

	Per Cent above Median			Chi-Squares Significant at the .05 Level		
	Under-achievers N=86	Normal-achievers N=81	Over-achievers N=98	U-N	U-O	N-O
U key	59.3	51.9	36.7	—	9.37	4.11
N key	50.0	49.4	39.8	—	—	—
O key	38.4	49.4	56.1	—	5.78	—
O-U score	41.9	45.7	65.3	—	10.16	6.94

cantly to the present predictors. However, the wide variation in the nature of the items comprising the key points to the difficulty of identifying variables by this approach.

Original impetus for the inclusion of the Strong in the present diagnostic battery came from its earlier apparent support of a hypothesis developed elsewhere (8, 9). Stated briefly, it was hypothesized that the extent to which behavior favorable to high grades will persist at the college level will be a function of the degree to which certain moral and cultural values have been internalized—i.e., positive deviation from predicted scores will be directly related to the phenomena variously labelled "superego," "conscience," "moral fiber," "goodness," etc.

Table 2 shows that in V, the so-called "goodness" group, when the samples are combined, on no scale do half of the U's exceed the median score. Correspondingly, more than half of the O's exceed the median on all Group V scales. Further, on three of the scales, the O's exceed the U's in all five samples. These findings, along with the incidence of statistical significances, indicate a relationship of Group V scales to academic achievement. This relationship was also found by Melville and Frederiksen (5) and by Morgan (6). Though the results are in the direction which the hypothesis would predict, it is still difficult to gauge the amount of support given to it. The instrument is empirical and any argument of support for the hypothesis must obviously involve a certain amount of tenuous reasoning.

The earlier finding that O's obtain lower M-F scores had been viewed as possible support for the hypothesis. This finding was corroborated by results yielded by additional samples. But, again, the difficulties in interpretation outlined above prevail.

The Group V scales merit special attention because of their possible measurement of a variable hypothesized to be related to achievement. Nevertheless, other occupational groups show similar discriminatory ability. In addition to Group V, O's score highest on occupational Groups I and X and also on the scales for Musician and C.P.A. Overachievers score lowest on occupational groups IV and IX. These findings seemingly do not bear on the authors' present hypothesis. It does seem, however, that high scores on occupations requiring extensive academic training are positively related to achievement.

Considerable agreement is found between our results and those of Melville and Frederiksen (5). The lack of greater agreement (especially on groups II and IV) may be due to differences in subjects. Melville and Frederiksen tested freshmen engineering students whereas the bulk of the present study's subjects were liberal arts students. Perhaps there are some personality or interest factors related to general academic achievement and others related to specific achievement.

Kendall (4) and Ostrom (7) found a positive relationship between achievement and O-L scores. The differences in design between these studies and the present one render direct comparisons difficult. Nevertheless, the results of the present study are in the same direction as those obtained by these authors.

Common among educators is the assumption that many students fail to achieve because of a disparity between occupational aims and measured interests. The results of this study fail to show that achievement is related to the congruency of occupational choice and scale scores. Morgan (6), using somewhat different criteria of achievement.



found a negative relationship between such congruency and achievement.

A comparison among the experimental groups in the combined samples yields 23 differences which are significant at less than the .05 level of confidence. Yet, none of these differences is obtained between U's and N's. Further, the empirical achievement keys, though able to distinguish the O's from both other groups, are unable to separate the U's from the N's. This result may be: (a) an artifact produced by the instrument; (b) due to the curvilinear nature of the variables related to achievement; or (c) produced by the academic structure which places a premium on overachievement while eliminating the extreme underachievers.

The suggestion of a curvilinear relationship between achievement and some variables has important implications for experimental design.<sup>4</sup> A large portion of published findings in this area is based on two contrasted achievement groups. This implicit assumption of linearity may be unjustified.

### Summary

The Strong Vocational Interest Bank for Men (SVIB) was administered to three groups of subjects (designated as under-achievers, normalachievers and overachievers) who were equated for general predicted score, but who differed in academic achievement. The blanks were scored for all occupations and comparisons among the groups were made by means of the chi-square technique. Empirical scoring keys were developed from two samples and cross-validated against three other samples.

1. The incidence of significant results obtained with both the occupational scales and the empirical keys was viewed as a demonstration of some relationship between achievement status and response to the Strong items.

2. The discriminatory ability of the Group V scales was regarded as lending possible support to the hypothesis that deviation from predicted grades is associated with a variable described as acceptance of or conformity to certain cultural values.

3. In general, the Strong does not seem highly appropriate for the measurement of the

<sup>4</sup> The author's results with the Rorschach (9) also indicate a curvilinear relationship similar to that presented here.

theoretical variable specified in the hypothesis.

4. Congruency between stated occupational aims and interests as measured by the Strong does not appear to be related to academic achievement.

5. Scale scores do not show a linear relationship with achievement; the overachievers appear to be the discrete group.

Received October 5, 1953.

### References

1. Crawford, A. B. and Burnham, P. S. *Forecasting college achievement*. New Haven: Yale Univ. Press, 1946.
2. Garrett, H. F. A review and interpretation of investigations of factors related to scholastic success in colleges of arts and sciences and to teachers' colleges. *J. exp. Educ.*, 1949-1950, 18, 91-138.
3. Harris, D. Factors affecting college grades: a review of the literature, 1930-1937. *Psychol. Bull.*, 1940, 37, 125-166.
4. Kendall, W. E. The occupational level scale of the Strong Vocational Interest Blank for Men. *J. appl. Psychol.*, 1947, 31, 283-287.
5. Melville, S. D. and Frederiksen, N. Achievement of freshman engineering students and the Strong Vocational Interest Blank. *J. appl. Psychol.*, 1952, 36, 169-173.
6. Morgan, H. H. A psychometric comparison of achieving and nonachieving college students of high ability. *J. consult. Psychol.*, 1952, 16, 292-298.
7. Ostrom, S. R. The OL key of the Strong Vocational Interest Blank for Men and scholastic success at college freshmen level. *J. appl. Psychol.*, 1949, 33, 51-54.
8. Rust, R. M. and Ryan, F. J. The relationship of some Rorschach variables to academic behavior. *J. Pers.*, 1953, 21, 441-456.
9. Ryan, F. J. Personality differences between under- and over-achievers in college. Ph.D. thesis, 1951, Columbia University. University Microfilms, Ann Arbor, Mich., Publ. No. 2857.
10. Strong, E. K., Jr. *Vocational interests of men and women*. Stanford University: Stanford Univ. Press, 1943.
11. Travers, R. N. W. Significant research on the prediction of academic success. In Donahue, W. T., Coombs, C. H., and Travers, R. N. W. (Ed.). *The measurement of student adjustment and achievement*. Ann Arbor: Univ. of Michigan Press, 1949. Pp. 147-190.
12. Young, C. W. and Estabrooks, G. H. *Scale for Measuring Studiousness by means of the Strong Vocational Interest Blank for Men*. Stanford University: Stanford Univ. Press, 1936.
13. Young, C. W. and Estabrooks, G. H. Report on the Young-Estabrooks Studiousness Scale for use with the Strong Vocational Interest Blank for Men. *J. educ. Psychol.*, 1937, 28, 176-187.

## Long-Term Validity of the Strong Interest Test in Two Subcultures

Charles McArthur

*Department of Hygiene, Harvard University \**

Surprisingly few long-term follow-ups have been made on the Strong Vocational Interest Blank, when one considers that the test has now been in use two decades. The largest studies are nine and ten year follow-ups reported in Strong's original volume (21), which are later supplemented by a twenty year report (22) on the same group. Unfortunately, as Super remarks (24), "the data are not so organized as to show what percentage of men entered and remained in fields in which they made A, B +, or lower scores." Instead, Strong (21) adduces support of four rather indirect propositions:

1. Men continuing in occupation A obtain a higher interest score in A than in any other occupation.
2. Men continuing in occupation A obtain a higher interest score in A than other men entering other occupations.
3. Men continuing in occupation A obtain higher scores in A than men who change from A to another occupation.
4. Men changing from occupation A to occupation B score higher in B prior to the change than in any other occupation, including A.

A special twenty-year follow-up by Strong (23) dealt with medical interests only but was reported in a more direct manner. Of 108 Stanford alumni who were physicians twenty years after testing, Strong reports that 70 had A ratings on the Physician scale in their undergraduate tests and 14 received a rating of B +. In all, then, 78% of these men who made careers as doctors had had a "high" physician score when tested in college.

### Procedure

*The Sample.* A series of 63 participants in the Study of Adult Development (then known as the Grant Study) were given the Strong Vo-

\*From the Study of Adult Development (the Grant Study), Department of Hygiene, Harvard University.

cational Interest Blank by Dr. F. L. Wells in the academic year 1939-1940. These young men were part of a longer series selected for interdisciplinary long-term study on the basis of their apparent "normality." All were at the time sophomores in Harvard College. Heath (8) has described the original program in detail.

This Study probably has the lowest rate of drop-outs of any existing longitudinal program. Of the 63 men given the Strong in 1939, only one has requested to be excused from further participation; all the rest are in close touch. It happens that the drop-out can be used in numerical summaries, since his occupation is known from perfectly public sources. Two men were lost during World War II, however.

We have, then, 61 cases on which to test the predictive power of the Strong over a fourteen year interval, from 1939 to 1953.

*SVIB as a Predictor.* How well did the Strong taken in college predict the occupations of these men fourteen years later? The basic detailed data for answering this question are given in Table 1.<sup>1</sup> Reported in Table 1 is the current job-title and the name of the Strong scale regarded as falling nearest to that job-title. Most selections are self-explanatory. One was semi-empirical; there being no scale for applied economists, it turned out that Office Man often came nearest. Disguises occur but only in the form of generalizing the job-title to make it less individually identifiable. The last two columns of Table 1 are mildly subjective evaluations by the investigator. It seemed necessary to specify whether or not a scale offered a "Direct" or an "Indirect" measure of interest in the occupation entered. The indirect measures are often no fair test at all, yet a counselor might in practice be forced to make just this sort of inference (e.g., using the Author-Journalist scale to assess the advisability of teaching Drama) for the lack of other evidence. In the last column, an assessment of the correctness of prediction is made in terms of "Good Hits," "Poor Hits," and "Clean Misses." The definitions of these terms are implicit in the claims made by Strong; he

<sup>1</sup>To reduce printing costs, Table 1 has been deposited with the American Documentation Institute. Order Document No. 4325 from the ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress, Washington 25, D. C., remitting in advance \$1.25 for 35 mm. microfilm or \$1.25 for 6 x 8 in. photocopies. Make checks payable to Chief, Photoduplication Service, Library of Congress.

Table 2

Fourteen Year Validation: Strong Vocational Interest Blank

Validity	Direct	Indirect	Total
Good Hit	22	5	27
Poor Hit	7	5	12
Clean Miss	14	7	21
Total	43	17	60

feels that a good hit may be counted when a man enters an occupation for which he scored A or which had the 1st, 2nd, or 3rd highest ranking score on his test. Less credence is given to a B + score when it is outranked by many others, yet such scores are usually regarded as "worth some consideration" in counseling. They are here called "Poor Hits." Anything below these criteria is taken to be a "Clean Miss."

Sixty cases could be used for validation, one man (No. 63) being in an occupation for which no scoring scale seemed even indirectly pertinent. It becomes apparent by inspection of Table 2 that some accuracy is lost through the necessity of using indirect measures. The fairest evaluation of the Strong's predictive power may be had from the 43 men whose occupations can be directly tested. Of these, only one-third are Clean Misses. Just half were hit well.

These figures are slightly lower than those given by Strong in his follow-up of medical interests. There, about one out of four tests turned out to be complete misses. Yet one must remain pleased with an instrument that under "blind conditions" (these tests were all unscored until 1952) predicts future behavior even half the time.

### Strong's First Proposition

Had a counselor used these tests, in 1939, to suggest to the boys their likeliest future vocation, he would have been downright misleading only once in every three attempts. Yet even the "good" tests would have presented him with a grave difficulty: the tests containing accurate predictions also contain too many "extraneous solutions." Like a mathematician solving a cubic equation, the counselor must enter the problem with the

expectation that not all the answers offered will be real and pertinent.

Whatever its letter rating, the scale most pertinent to future choice of occupation ranked anywhere from 1st to 33rd highest out of the 44 scales for which each test was scored. The median rank of the most pertinent scale was 5th. That means that the counselor using these tests could have expected, on the average, four "extraneous solutions" with higher-ranking scores than the true solution. It is, of course, true that the "extraneous" quality of certain high scores is obvious: few would counsel a tone-deaf boy to be a musician.

Strong (21) states that "a college student who continues ten years in the same occupation enters an occupation in which he ranks second or third best." Like our group as a whole, our men who continued in the same occupation (not considering interruption by the war) entered occupations in which, on the median, they ranked fifth best. Once again, our figures are slightly less impressive than Strong's. It is certainly not true that among our cases men "continuing in occupation A obtain a higher interest score in A than in any other occupation."

### Strong's Second Proposition

The proposition that men engaged in an occupation score higher on that occupational

Table 3

Testing Strong's Second Proposition

Occupation	Average Score of Men Engaged in Occupation	Average Score of All Other Men
Physician (N = 12)	42.3	32.8
Lawyer (N = 11)	40.6	30.5
Public Administrator (N = 5)	45.8	39.6
Engineer (N = 4)	53.8	30.1
Chemist (N = 3)	45.0	33.2
Minister (N = 2)	44.0	29.2



scale than all other men is well supported by our data. That is, doctors outscore controls on the Physician scale, lawyers outscore controls on the Law scale, etc. (Controls are simply all the rest of the 61 cases.) This is true for every directly scaled occupation that occurs more than once.

Strong's second proposition seems to be valid.

### Strong's Last Two Propositions

Seventeen of our sixty men have made changes in occupation other than shifts enforced by entering the armed services. Often, these men abandoned two or more vocations before settling on the job they are engaged in today. Strong's follow-up data showed that men who abandoned an occupation were likely to possess lower scores on that occupational scale than the scores made by men who continued on the job.

Table 4 tests that proposition in our own figures. Strong found that rule to hold "except for the records of two individuals," while we, except for one instance of tie, find it to be entirely so.

Another generalization Strong offers about men who change vocational fields is that they

Table 4  
Testing Strong's Third Proposition

Occupational Scale	Men Continuing		Men Leaving	
	Mean Score	N	Mean Score	N
Physician	42.3	12	35.5	2
Lawyer	40.6	11	33.0	3
Public Admin.	45.8	5	42.5	4
Author-Journalist (Teaching)	49.7	3	33.0	3
Engineer	53.8	4	34.0	2
Office Man	44.0	2	35.0	3
Production Mgr.	30.0	2	21.0	2
Pres. Mfg. Co.	25.3	3	34.0	1
Physicist	42.5	2	34.0	2
Chemist	45.0	3	18.0	1
Author-Journalist (Writing)	54.0	1	32.5	2
Minister	44.0	2	35.0	1
Salesman	42.0	1	30.0	1
Senior C.P.A.	43.0	1	43.0	1

will proceed from a field in which they have a low score into a field in which they score high. That was true of 9 of our changeable men, 7 men going contrary to their tests and entering new jobs for which their test scores were lower. (One man changed between jobs with identical scores.) These figures run faintly in the right direction, probably looking even less convincing than the data from which Strong felt that proposition 4 was "almost but not quite sustained."

### Contentment in Occupation

As Strong has pointed out (23), "the validity of an interest test should be measured in terms of satisfaction" but for this "there is no satisfactory measure." The Study of Adult Development has accumulated much data on expressed satisfaction and dissatisfaction with occupational choice, through the use of annual questionnaires. Even as we heed the force of Murray's (16) warning that one must draw conclusions from inferred sentiments, not from expressed sentiments, we may ask some operational questions about the relations between expressed satisfaction in 1953 and the interest score obtained 14 years previously.

The 1953 questionnaires were still coming in when this was written. Of the 60 men in whom we are interested, 37 had returned their questionnaires. There was, as a matter of fact, some tendency for the men engaged in occupations for which they possessed a favorable Strong score to return their questionnaires early! (Three-quarters of them had done so, as against half the men with lower scores. For this Fisher's "p" comes out .09. This is not so trivial an indication as it may appear; the Study staff has long been aware that among people who are hardest to hear from are those who have a sense of not having succeeded.

Several 1953 questions were pertinent to an inferrable sentiment of job satisfaction. They may be abbreviated as: (a) Are you contemplating a change in the near future? What considerations entered this? (b) To what extent has the job produced strains? (c) What special events have occurred in the last year? (d) What is your outlook on your

Table 5

## Job Satisfaction and Strong Score

Score	Appar- ently Happy	Express Discon- tent	Total
"A" on Strong scale	14	3	17
Lower scores on Strong scale	10	10	20
Total	24	13	37

personal future? And what is the principal basis of this? Not rarely, a participant makes use of the backs of the questionnaire pages to write us a letter in which discussions of job problems may be found.

There were thirteen men, in all, who showed *some* evidence of discontent, in answer to one or another of the questions. These thirteen, who are "less than completely happy" about their jobs, include disproportionately few who scored A on the Strong.

Table 5 gives the figures. Fisher's "p" comes out less than .05.

The question about job strains is the only one of those contributing to this general result that itself approaches significance. Though the figures in Table 6 are not impressive, "p" comes down to .08.

The contributions of the other questions, though in the predicted direction, are too small in numbers to reach significance. (An example: men now occupying the lower-rated

Table 6

## Job Strains and Strong Score

Rating	No Strains Reported	Reported Strains	Total
A or B+ Strong Rating	21	3	24
Lower Strong Rating	8	5	13
Total	29	8	37

occupations are twice as frequently contemplating a change.)

## Other Evidence

These findings, though not so favorable to the test as Strong's results, nonetheless suggest that the test has its usefulness. Furthermore, someone familiar with the Study participants cannot read through Table 1 without acquiring some feeling that, however inaccurate its predictions of *behavior*, the test is measuring *interests*. There is the evidence, for example, of the correlated pair of scores: Lawyer and Public Administrator. Some men enter the law because they have politics in mind. Cases 20, 25, and 27 are examples. In case 27, the Public Administrator score matches that for Lawyer. In case 25, the Lawyer score is low; the choice of lawyer would seem to have been contraindicated. That would have been correct. Case 25 escapes being one of our dramatically unhappy group only because the practice of law is rationalized as a means to a political end. The Strong has measured the relative interest in law and politics quite accurately. Indeed, the suggestion of power motives given by the Strong is more than borne out by projective tests. (Case 24 is in sharp contrast. Though actually working for the government, this man is not interested in politics. That is what his Strong scores fourteen years ago predicted.) Some indication of the injustice of "occupations entered" as a criterion of interest may be had from case 20. In the table, this man is reported as a lawyer and his lowish score on that scale makes him count in the validation as a "Poor Hit." Yet he, too, intends to use law as a stepping-stone into politics, a fact that was not shown in the table, since circumstances have prevented his carrying out his plans. His score on Public Administrator is an A. That is also the scale on which he ranks first.

One is impressed by the logic underlying the relative efficacy of the test in predicting well or poorly certain occupational choices. Engineers, ministers, and teachers seem to be highly predictable; all three are likely to choose their vocation in response to an inner "call." By contrast, men who are in their

own business (which, for all three under that heading in Table 1, means an "externally prescribed" choice) the Strong simply does not predict. Another way of saying these facts would be to assume that the Strong tested interest and that the difference in prediction represented differences in the importance of interest as a factor in various sorts of career choice. The very patterning of the failures of the test therefore confirms its validity as a measure of interest!

### Private and Public School Results

Suppose we explore the consequences of postulating that the Strong does measure interests. We infer that the test will predict future job-choices only for those men who (consciously or unconsciously) give weight to their own interests when they choose a career. For men who do not follow their interest, the test will not predict. We therefore expect the Strong's "validity" to vary between groups known to take their own interests more or less seriously. A major instance of such a prediction is provided by our tests from men who prepared for Harvard at public and private secondary schools.

The public school boy has usually been raised in the "American success culture," described by many anthropologists (1, 2, 4, 10, 15). His parents' efforts focussed on preparing the boy for future vocational achievement. Job choice has been for him a vital matter; his future self-estimate will hinge on his job-title and on how well he does within his occupational field. As one Study participant explained it, "I have satisfied myself as to my ability to compete successfully with most of my contemporaries."

The private school boy will often have been reared in a variant orientation, ably described by Florence Kluckhohn (10), where child-rearing was intended to perpetuate in him a "preferred personality." Occupational role will have been subordinated to family social patterns. In our 1953 questionnaires eleven private school boys put family interest or personal breadth ahead of achievement values when discussing their "personal future." As Kluckhohn so nicely phrased it, the contrast is be-

tween two subcultures, one emphasizing a "Doing," the other a "Being," orientation.

One consequence of this subcultural contrast is a difference in the importance assigned to interests when men make their vocational choice. In the "success culture" a son is expected to surpass (therefore often bypass) his father's occupation. Choosing a job is for him a vital matter, the more so because the choice is so greatly "up to him." So much hinges on his making a "right" choice, calculated to yield maximal success, that he will often consult his own interest pattern, either introspectively or with formal aid from a vocational counselor. By contrast, the purest case of the upper class variant is a man whose permitted choices are limited to three: trustee, lawyer or doctor. Patricia Smith (20) described the sanctions that suppress other alternatives. (The Study has witnessed dramatic conflicts within upper class men when personal "calls" gave way before the pressure of tradition.) While the average private school boy is not subjected to so focal a pressure, he will nevertheless possess values reinforcing the tangible demand that he join his father or uncle in *The Business* and the intangible expectation that he will first of all be the Right Sort. As one participant wrote, "As near as I can tell I have those (personal) qualities in some small measure, so I think it foolish to spend time thinking about my future."

If all this is true, we arrive at the prediction that interests will matter less and therefore the Strong will be less valid when applied to the behavior of private school boys. Table 7 shows this to be the case. Chi square suggests  $p$  less than .05; if we combine cells

Table 7

Validity of Strong Test Applied to Public and Private School Boys

Validity	Public	Private	Total
Good Hit	19	8	27
Poor Hit	4	8	12
Clean Miss	8	13	21
Total	31	29	60



(avoiding the low cell and isolating the relation between public school attendance and "Good Hits"), we can apply Fisher's formula and arrive at  $p$  below .01. Our proposition seems well validated.

If we translate Table 7 into percentage, we discover that three-quarters of the public school tests gave some sort of "hit" on the occupation engaged in fourteen years after testing. That is exactly the figure reported by Strong (23) for his twenty-year follow-up. If, on the other hand, we try to apply the test to private school boys, our predictions will be useless almost half the time.

Splitting out the public school cases, we can try revalidating Strong's four propositions. Proposition 1 fares better: men engaged in occupation A still do not have "a higher interest score in A than in any other occupation" but the median rank of the pertinent scale is third, where formerly it was fifth. That is more consistent with Strong's claim, quoted earlier, that the occupation continued in will have ranked first, second, or third. Proposition 2 is no better for the public school group alone; that is because some occupations (engineer, chemist) attract high scores from public school, while others (lawyer, minister) attract higher scores from private school. At any rate, Proposition 2 was already verified sufficiently. Proposition 3 was already verified in every comparison, and so cannot be improved. There is one scale (Public Administrator) on which Proposition 3 is false for the private school group but true for the public school group. Proposition 4 is about equally valid in both groups.

### Discussion

This finding will raise various questions, some of which can be answered from our data. To forestall one, the "private school effect" cannot be explained in terms of income. It is true that the Strong is less accurate when applied to families receiving over sixteen thousand dollars a year, but this figure marks only the upper quartile of our income statistics, while the "private school effect" is visible at all income levels. For example, in the second income quartile, with income held reasonably constant, between four and six thousand dol-

lars, public school tests score good hits 75% of the time, private school tests only 40%. In all income quartiles that are adequately represented by public school cases, the proportion of misleading tests remains about 1 in 4; in all income quartiles that are adequately represented by private school cases, the proportion of misleading tests remains about 1 in 2.

These figures suggest that it is the fact of having attended private school (or of being reared in a subculture from which one is sent to a private school), rather than income, and somewhat independently of social class, that depressed the validity of the test. Several explanations suggest themselves. The most obvious would be that the Strong was validated against public school graduates. (Regional differences in patterns of secondary education would have led to this circumstance.) Next most obvious might be that attending private school is one of those "experiences affecting interests" that Super (24) warns us have been too little studied.

### Related Findings

The effects of private school mores on personality reported here are not isolated phenomena. Private school boys have previously been assumed to possess a special system of values, by scientists (10, 20, 25, 26), novelists (11, 12, 17) and deans (27). Empirical demonstrations show that their responses differ from those of public school boys on projective tests (13), especially with regard to the projection of need Achievement (14), the need that underlies the results reported here. What is said here of their attitudes to vocational success has long been known with regard to their attitude toward academic success (18) and the effect of this attitude on their grades has long been empirically demonstrated (5, 6, 19). Very much that is known about this topic remains unpublished.

It seems to the writer that psychologists in Eastern universities, by failing to report the public-private school differences in their data, are failing to record a fine "natural experiment" in the laws governing culture and personality. The New England private school boy is often that rarest of subjects in the

psychological laboratory: a member of one of America's geographically scattered upper classes (3). The Chicago group (1, 2, 7, 9, 25, 26) has done much to call our attention to differences between middle and lower class personalities. Are not differences between subcultural personalities in the middle and upper classes likely to be just as great?

### Summary and Conclusions

A fourteen-year follow-up was made of Strong Vocational Interest Blanks administered in 1939 to participants in the Study of Adult Development. The validity of the test as a predictor of occupational choice at first appeared to be slightly lower than that reported by Strong. Of Strong's four validation propositions, two were confirmed, one (that lawyers outscore non-lawyers on the Law scale, etc.) strikingly, the other (that lawyers obtain one of their best scores on the Law scale, etc.) less so. The median test offered four "extraneous" predictions.

It was possible to demonstrate a relation between conformity to choices commended by the test and future vocational happiness. Choosing a job for which one had (some years before) scored "A" also seemed to reduce the likelihood of developing fatigue, irritability or other symptoms of strain.

The proposition was offered that SVIB validly measured interests but that failure to predict what job a man would choose could be explained in terms of his making the choice on some basis other than interest. Certain case histories supported this idea as did the apparent pattern in occupations which the test predicted accurately and which it did not.

As a corollary of this proposition and on the basis of what has been learned elsewhere, it was predicted that the Strong would be applicable to boys who attended a public secondary school but less useful for boys who had prepared in a private preparatory institution. That was the case. The predictive validity of the test among the public school group was almost exactly that originally reported by Strong. Among private school boys, the test was, half of the time,

inapplicable. Further, Strong's first validation proposition was improved in the public school group, the median test record offering only two extraneous predictions.

The import of this finding may be read in one of two ways. If we assume the anthropological theories about the American middle and upper classes to be true, then this is a demonstration that "invalidity" in the Strong arises because interests do not determine choice rather than from failure of the test to measure interests. On the other hand, the implication that there may be a distinct psychology of the upper class is also pointed out.

From all this may be drawn the following conclusions:

1. The Strong has at least the validity claimed for it as a measure of interests.
2. Its most rigorous validation criterion will be the prediction of actual behavior, but even that criterion is met at least 1 time in 2.
3. We may regard as critical for understanding the use of the test Strong's (23) proposed "future calculations as to how much other factors, such as economic conditions, family pressures, etc. affect a man's occupational career." In this respect attention should be called to upper class variants of the American personality.

Further study of: (a) the effects of environmental press in conflict with interests measured by the Strong; and (b) the differences between public and private school personalities will be made from Study of Adult Development data.

Received September 21, 1953.

### References

1. Davis, A. *Social class influences upon personality*. Cambridge: Harvard University Press, 1948.
2. Davis, A. American status systems and the socialization of the child. In Kluckhohn, C. and Murray, H. A. (Eds.), *Personality in nature, society and culture*. New York: Alfred A. Knopf, 1949.
3. Goldschmidt, W. Social class in America: a critical review. *Amer. Anthropologist*, 1950, 52, 483-498.
4. Gorer, G. *The American people*. New York: Norton, 1948.

5. Harris, D. The relation to college grades of some factors other than intelligence. *Arch. Psychol.*, New York, 1931, 20, no. 131.
6. Harris, D. Factors affecting college grades; a review of the literature, 1930-1937. *Psychol. Bull.*, 1940, 37, 125-151.
7. Havighurst, R. J. and Taba, H. *Adolescent character and development*. New York: Wiley, 1949.
8. Heath, C. W. et al. *What people are*. Cambridge: Harvard University Press, 1945.
9. Hollingshead, A. B. *Elmtown's youth*. New York: John Wiley & Sons, 1949.
10. Kluckhohn, Florence R. Dominant and substitutive profiles of cultural orientations: their significance for the analysis of social stratification. *Social Forces*, 1950, 28, 376-393.
11. Marquand, J. P. *The late George Apley*. Boston: Little, Brown & Co., 1937.
12. Marquand, J. P. *Point of no return*. Boston: Little, Brown & Co., 1949.
13. McArthur, C. C. *Cultural values as determinants of imaginal productions*. Unpublished doctor's dissertation, Harvard University, 1951.
14. McArthur, C. C. The projection of need Achievement: a re-examination. *J. abnorm. soc. Psychol.*, 1953, 48, 532-536.
15. Mead, Margaret. Has the 'middle class' a future? *Survey Graphic*, 1942, 31, 64-67, 95.
16. Murray, H. A. and Morgan, Christiana. A clinical study of sentiments. *Genet. Psychol. Monogr.*, 1945, 32, 3-149, 153-311.
17. Phillips, J. *The second happiest day*. New York: Harper & Bros., 1953.
18. Sarason, S. B. and Mandler, G. Some correlates of test anxiety. *J. abnorm. soc. Psychol.*, 1952, 47, 810-817.
19. Seltzer, C. C. Academic success in college of public and private school students: freshman year at Harvard. *J. Psychol.*, 1948, 25, 419-431.
20. Smith, Patricia. *The problems of occupational adjustment for the upper class Boston man*. Unpublished honors thesis, Radcliffe College, 1950.
21. Strong, E. K., Jr. *Vocational interests of men and women*. Stanford University: Stanford University Press, 1943.
22. Strong, E. K., Jr. Interest scores while in college of occupations engaged in twenty years later. *Educ. psychol. Measmt.*, 1951, 11, 335-348.
23. Strong, E. K., Jr. Twenty year follow-up of medical interests. In Thurstone, L. L. (Ed.), *Applications of psychology*. New York: Harper & Bros., 1952.
24. Super, D. E. *Appraising vocational fitness*. New York: Harper & Bros., 1949.
25. Warner, W. L. *Social life of a modern community*. New Haven: Yale University Press, 1941.
26. Warner, W. L., Havighurst, R. T., and Loeb, M. L. *Who shall be educated?* New York: Harper & Bros., 1944.
27. "The College." In Reports of the president and treasurer of Harvard College, 1923-4. *Official register of Harvard University*, vol. 22, no. 5, February 24, 1925.



## Vocational Interests of Naval Aviation Cadets \*

Nathan Rosenberg and Carroll E. Izard

*The Tulane University*

In interviews with cadets who voluntarily withdraw from the Naval Air Training Program, an active dislike of flying was one of the most important expressed reasons for withdrawal (1). An attempt was made, therefore, to investigate the importance of interests as a correlate of success in Naval Aviation. This attempt was directed toward an examination of broad interest patterns of cadets through measurement of their vocational interests rather than dealing with specific interests in flying and the training program itself. Since questions about flying and the program are avoided in tests of vocational interests, such measures were considered more subtle, less subject to momentary fluctuations in attitudes that seem present in newly arrived cadets, and of greater psychological importance.

The *Kuder Preference Record, Vocational, Form B*, (3) was chosen to measure vocational interests since it is one of the interest questionnaires which has been most widely studied for validity. Form B was selected because it had been administered in World War II to a population of Air Force cadets. The writers feel that Navy and Air Force attrition samples differ in many important characteristics, and the proposed comparison will present definitive evidence with regard to measured vocational interests. Generalizations from World War II Air Force data are often made concerning the importance of many psychological characteristics for selection of pilots. If this Air Force population differs in important respects from other aviation populations, such generalizations should be tempered.

At the outset certain methodological considerations should be noted. It is reasonable

\* This article was presented as a report to the U. S. Naval School of Aviation Medicine, Pensacola, Florida, under ONR Project NR154-098. Opinions or conclusions contained in this report are those of the authors. They are not to be construed as necessarily reflecting the views or possessing the endorsement of the Navy Department.

to assume that certain vocational interest patterns may cause cadets to enter Naval Air Training. Once this pre-selection has operated, there may or may not be a relationship between interests and successful completion of training. That is, interests may cause entry into training but they may or may not be predictive of success after pre-selection has occurred. Thus, it is important to consider whether naval aviators possess distinguishing interests prior to entry into training.

Should selective drop-out during training occur, it is possible that interests operate as a post-selective device. This implies a correlation between interests and successful completion of training. This correlation is most adequately tested by a longitudinal approach in which entering cadets are tested and then followed through the program to identify successful and non-successful cases. A compromise to this longitudinal study is afforded by the cross-sectional approach. Entering cadets, non-successful cadets, and successful cadets are tested and their mean interest scores compared. Mean scores which differ *systematically* are interpreted as evidence for a correlation between interests and successful completion of Naval Air Training.

Should training, or factors operating during training, change the interests of entering cadets, the change might contaminate inferences regarding test validity. For example, maturation of cadet interests over an 18 month training period might well influence apparent test validity. In this report, selective drop-out during training is assumed to result from differences in interests between an attrition and successful group. The preceding considerations have been made so that appropriate safeguards will be followed in interpreting the results.

The following questions are considered in this report:

1. Do the vocational interests of entering Naval Aviation Cadets differ significantly

Table 1

Means and Standard Deviations for Kuder Interest Scores on Various Groups Considered

Kuder Interest Area	Entering Naval Aviation Cadets N=651		"Successful" Cadets N=137		(DOR) Voluntary withdrawals from Training N=137		World War II Air Force Cadets N=937		Kuder's Normative Population N=2667	
	M	SD	M	SD	M	SD	M	SD	M	SD
Mechanical	81.5	17.8	85.8	16.3	73.3	20.9	86.0	15.6	78.6	22.8
Computational	33.9	11.1	33.2	11.3	33.3	12.6	33.2	9.3	35.3	10.6
Scientific	70.7	14.8	68.4	14.5	61.1	16.4	67.6	12.6	64.0	15.5
Persuasive	71.7	18.8	73.9	19.9	82.3	20.4	68.4	16.8	74.4	20.6
Artistic	50.9	13.4	53.1	14.4	50.4	16.0	49.3	13.3	46.1	13.6
Literary	38.6	13.6	35.5	11.1	41.1	14.4	46.4	13.3	47.8	15.1
Musical	19.6	9.5	17.2	8.6	19.7	9.5	19.0	9.0	16.6	9.6
Social Service	66.4	17.0	65.7	16.3	69.3	16.5	63.7	14.3	73.7	17.5
Clerical	41.0	11.6	42.7	12.7	44.9	14.1	46.4	12.1	52.1	13.5

from an unselected vocational group, namely the norm group found in the test manual for the *Kuder Preference Record*?

2. Do the vocational interests of successful Naval Aviation Cadets differ significantly from an attrition population of cadets who withdraw at their own request?

3. Do the vocational interests of present-day Naval Aviation Cadets differ significantly from a wartime Air Force population of cadets?

### Procedure

*Samples Used.* 1. Entering classes 3-53 through 10-53 and classes 16-53 through 23-53 were tested. A total of 16 classes consisting of 651 subjects were included in this group. By the completion of Naval Air Training, about 15 per cent of an entering class will have withdrawn voluntarily. Attrition from all other causes generally averages to about this same percentage; thus total attrition averages about 30 per cent.

2. The successful group consisted of 137 cadets who were tested at Corry Field, approximately nine months after entry into training. Based upon previous experience, it is estimated that over 90 per cent of these subjects will graduate.

3. A total of 137 DOR cases (Dropped at Own Request) were tested, as many as administratively possible during the period from about 1 January through 1 June 1953.

4. The norm group consisted of 2,667 adult men engaged in diversified occupations, obtained from the manual for the *Kuder Preference Record*.

5. From published Air Force data, results were available for 937 wartime cadets, 721 of whom

graduated primary training and 216 of whom were eliminated (2).

### Results

Table 1 presents a summary of means and standard deviations for the nine interest areas measured on the groups considered. Table 2 shows critical ratios testing the significance of the differences in mean interest scores for the groups compared.

*Comparison of Entering Cadets' Interests to Kuder's Norm Group.* The norm group consists of "2,667 adult men engaged in occupations, with each major occupational group weighted in proportion to its occurrence in the general population (with the exception of unskilled and semi-skilled workers)" (3).

On the average (Tables 1 and 2), entering cadets possess significantly different interests from those found for Kuder's norm group in all nine interest areas measured. Entering cadets are relatively more interested in scientific, artistic, musical, and mechanical activities and relatively less interested in clerical, literary, social service, persuasive, and computational activities.

Another method of evaluating the difference in interests between the two groups is gauged by the following procedure. The mean interest scores for entering cadets and the norm group are placed on the distribution of scores for the norm and percentile ranks

Table 2

Critical Ratios Testing Significance of Difference in Mean Kuder Interest Scores †

	Mec	Com	Sci	Per	Art	Lit	Mus	SS	Cle
A. <i>Entering Cadets versus Norm Group</i>									
C.R.	3.42**	2.81**	10.22**	3.20**	8.08**	14.97**	7.12**	9.76**	21.38**
B. <i>Successful Cadets versus Voluntary Withdrawals (DOR)</i>									
C.R.	5.52**	0.06	3.92**	3.46**	1.43	3.62**	2.32*	1.82	1.33
C. <i>Entering Cadets versus World War II Air Force Cadets</i>									
C.R.	5.27**	1.38	4.32**	3.61**	2.37*	11.25**	1.34	3.28**	9.02**

\* Significant at the 5% level of confidence.

\*\* Significant at the 1% level of confidence.

† Mec—Mechanical; Com—Computational; Sci—Scientific; Per—Persuasive; Art—Artistic; Lit—Literary; Mus—Musical; SS—Social Service; Cle—Clerical.

obtained. Percentile ranks obtained by this procedure are presented in Table 3.

In a perfectly normal distribution, the mean interest scores for the norm group would all lie at percentile rank of 50, the median score. Deviations from a percentile rank of 50 for the norm group suggest the direction and degree of skewness for the norm distribution. Since all percentile ranks for the norm group appear fairly close to 50, the skewness, if significant, would not appear pronounced. Inspection of the norm distribution for mechanical interest, where the mean score approximates a percentile rank of 45, suggests that mechanical interest scores are slightly skewed toward the high end of the distribution. This explains an ap-

parent contradiction whereby entering cadets show a mean mechanical interest score equivalent to a percentile rank of 50 on the norm distribution and, at the same time, show significantly greater interest in the mechanical area than the norm.

The extremity of the differences between entering cadets and the norm group is emphasized for the clerical, literary, and social service areas. Entering cadets seem pre-selected with respect to a relative dislike for activities of reading or writing (literary), routine filing or secretarial work (clerical), and activities which contribute to the welfare of people (social service). To a lesser extent, they are pre-selected with respect to a relative liking for activities of the scientific, artistic, and musical interest areas.

Since the norm group is presumably older than the cadet group, it may not be concluded that these differences are all characteristic of Naval Cadets as a vocational group. Some of the differences could reflect changes in interest characteristic of an older age group. Furthermore, cadets undoubtedly represent a population with more education than do the norm group. Thus some of the differences in interests could be a reflection of educational level which distinguishes the two groups, aside from vocational selection. When these factors are better controlled, it will be possible to isolate which of the interest areas reflect those characteristics of a vocational group and not those for age or educational groupings.

Table 3

Percentile Ranks of Interests for Entering Naval Cadets and Norm Group

Interest Area	Entering Naval Cadets	Norm Group	Difference Between Entering and Norm Group
Mechanical	50	45	+ 5
Computational	45	49	- 4
Scientific	67	50	+17
Persuasive	48	52	- 4
Artistic	65	54	+11
Literary	30	54	-24
Musical	65	57	+ 8
Social Service	30	50	-20
Clerical	20	52	-32



It would seem reasonable that a preference for the scientific area would be the one area most likely to be truly characteristic of Naval Aviators as opposed to vocationally unselected groups.

*Comparison of Successful Cadets to Voluntary Withdrawals (DOR).* Differences in interests between the above two groups suggest the possible usefulness of the *Kuder Preference Record* as a predictor of DOR attrition. As can be noted from Tables 1 and 2, successful cadets are significantly more interested in mechanical and scientific activities than DOR's. They are significantly less interested in persuasive, literary, and musical interests than DOR's.

From these results, the interest picture for the successful cadet is an individual who has a positive attraction toward activities which involve the use of tools and machinery; he also likes abstract and theoretical activities of a scientific nature. The DOR appears to be an individual who is more interested in activities which involve convincing people (persuasive), reading or writing, and appreciation or participation in musical activities; he is less attracted by mechanical and scientific activities.

In this connection, it should be recalled that entering Naval Aviation Cadets are selected with respect to mechanical aptitude since cadets with very low Mechanical Comprehension Test scores are not admitted to the training program. These data indicate that mechanical interest, aside from mechanical aptitude, is important for successful completion of the Naval Air Training Program. Further study will be made to evaluate the improved prediction of DOR attrition when aptitudes and interests are both considered.

*Comparison of Entering Cadets' Interests with an Air Force Population.* Critical ratios (Table 2) reveal some important differences in interest between the above two groups. Entering cadets' interests differ significantly from the Air Force entering cadet population in all areas with the exception of computational and musical activities. The differences between the two groups are particularly pronounced for the literary, clerical and mechanical areas.

Inspection of the mean scores for the Air Force eliminees from training reveals that differences in interests between the two attrition groups are considerable.<sup>1</sup> The Naval Cadet who withdraws voluntarily shows essentially a different interest pattern from the Air Force cadet who was eliminated from training during World War II. The reasons for this difference are not very clear, aside from motivation present during World War II which is not so pronounced today. However, the important fact is that these two populations are different—at least with respect to interests. Thus if a test did not show validity on the Air Force population of World War II, this does not necessarily preclude its being valid for present day Naval Aviators. The attempt to use the Kuder in this study was undertaken despite Air Force data which showed it to be invalid for predicting pass-fail during World War II (2).

### Discussion

It will be recalled that successful cadets as compared with DOR groups possess higher mean interest scores for mechanical and scientific areas and lower for the persuasive, literary, and musical areas. The mean interest scores for entering Naval Aviation Cadets lie between those found for successful and DOR groups for mechanical, literary, and musical interest areas (Table 1). These findings for the entering group are consistent with the assumption that selective drop-out from training caused the significant differences noted between successful and DOR groups. However, the scientific interest area deserves special comment since the mean scientific interest score for the successful group is 68.4, for the DOR group 61.1, but for entering cadets 70.7. Although entering cadets are more like the successful than the DOR, for a definite trend to be present, the mean interest scores

<sup>1</sup> Mean interest scores for Air Force eliminees from training differ by only a small fraction of a point from those for the Air Force graduates, with the exception of artistic and social service interests. Eliminees are about 2.0 and 2.5 points higher and lower in these two interest areas respectively. Thus, interest comparisons may be made directly to the total Air Force population means with little loss of accuracy as compared to the eliminees from this population.

for entering cadets should lie between the means for successful and DOR groups. The same reasoning applies for persuasive interest where the mean score for the entering group does not lie between the DOR and successful groups.

It is possible that entering cadets tend to over-rate their interest in scientific activities. Having just reported to Naval Air Training, it is conceivable that they would tend to rate themselves higher in scientific interest merely because they feel they *should* be high in this interest.

Since further "cross-validation" will be applied to these data in any case, an empirical check will be made for those interest areas apparently important for successful completion of Naval Air Training. Based on the differences between successful and attrition cases, weights will be given to the interests that distinguish the two groups. From these weights, predictions of pass or DOR attrition will be made for entering cadets. In time, cadets who actually voluntarily withdraw and those who succeed will be determined. These results will be checked against the predictions made, and the actual utility of the *Kuder Preference Record* for predicting DOR cases will be ascertained. From the results presented in this report, it seems very likely that the measured vocational interests of entering cadets will predict DOR attrition significantly greater than chance expectation.

Summary

The vocational interests of cadets would seem important for successful completion of Naval Air Training. Therefore, the *Kuder Preference Record*, a measure of relative preference for nine broad vocational interest areas, was administered to 651 entering Naval Aviation Cadets; 137 DOR attrition cases (voluntary withdrawals from training) and 137 "successful" cadets. The successful cadets were tested near completion of their basic training; from previous experience it is estimated that over 90 per cent of these cadets will graduate.

Results indicate:

1. Entering cadets show significantly more interest in scientific, artistic, musical, and mechanical activities than a vocationally unselected population. They are less interested in clerical, literary, social service, persuasive, and computational activities.

2. Successful cadets are relatively more interested in mechanical and scientific activities as compared to a group who withdraw from training at their own request. They are less interested in persuasive, literary, and musical activities than the voluntary withdrawal cases.

3. The voluntary withdrawal group shows an essentially different interest pattern than the group eliminated from training in the Air Force during World War II.

It is concluded:

1. Entering cadets have interest patterns which are different from those found for a vocationally unselected group. Some of these distinguishing interests may arise because of cadets' age or educational level rather than choice of Naval Aviation as a vocation. The factor of selection screening, as well as self-selection on the basis of interests, may have partially determined these interest patterns.

2. The *Kuder Preference Record* shows promise of validity for predicting DOR attrition. The mechanical, scientific, persuasive, literary, and musical interest keys appear the most important for this purpose.

3. Some psychological tests which failed to predict attrition for World War II Air Force cadets may show validity for present day Naval Aviators.

Received October 23, 1953.

References

1. Bair, J. T. and Ambler, R. K. *Expressed reasons and background characteristics for Naval Aviation Cadets withdrawing voluntarily during January 1953.* U. S. Naval School of Aviation Medicine. Special Report—Attrition Report No. 6, February 1953.

2. Guilford, J. P. and Lacey, J. I. (Eds.) *Printed Classification Tests.* Report Number 5, Army Air Force Aviation Psychology Program Research Reports, 1947.

3. Kuder, G. F. *Revised Manual for the Kuder Preference Record.* Chicago: Science Research Associates, 1946.



## Coding the Kuder Preference Record—Vocational \*

Robert Callis, William C. Engram, and John F. McGowan

*University Counseling Bureau, University of Missouri*

Underlying the use of vocational interest tests in vocational planning is the assumption that if a person's interests are similar to the interest of people in occupational groups who have experienced a high degree of satisfaction in their work, he will derive most satisfaction doing the same or similar kind of work. That is, if a person's interests in "common-everyday things" are most similar to, say, engineers, there is a high probability that he will derive more satisfaction from work as an engineer or some closely related occupation than he would from other occupations. There has been considerable research to substantiate this proposition. In order, then, for a counselor to be effective in the interpretation of his client's interests as measured by tests, he needs some sort of guide to aid him in giving the client a comparison of his interests with that of various occupational groups.

This paper presents a guide for the counselor to use in interpreting the individual profile of the *Kuder Preference Record—Vocational* (Kuder PR-V). In order to facilitate a meaningful interpretation of the test data to the client, it is often better for the counselor to speak in terms of several occupational fields in addition to descriptive terms (such as mechanical, artistic, persuasive), the meaning of which is often vague to the client. There is, therefore, a need for a grouping of occupations based on real test data which the counselor may feel confident in using.

Kuder (12) grouped specific occupations under the various scale headings of his inventory. However, many counselors have been reluctant to interpret the client's profile in terms of Kuder's groupings of occupations because many of the groupings were not sup-

ported by empirical data. During the past few years Kuder and many others have reported a considerable amount of empirical data about various occupational groups which can be used to group occupations according to interest test profiles. However, some of the discrepancies between Kuder's grouping (12, Table 1) and his empirical data (12, Tables 2 and 3) are rather striking. Wiener (20) cited as an example of one of these discrepancies Kuder's "39" listing (Scientific and Clerical interests) as including the occupation of pharmacist. Looking at actual test results for a group of "pharmacists and drug store managers," however, one sees a significant elevation on scale 3 (Scientific) and only an average score on scale 9 (Clerical).

We find, as another example, that Kuder lists "Author; editor; reporter" under the categories of "4" (Persuasive), "6" (Literary), "36" (Scientific-Literary), "46" (Persuasive-Literary), "67" (Literary-Musical) and "68" (Literary-Social Service). From empirical research, Mathewson and Herbert (14) found that only the "67" category was the pattern for their group of 113 author-journalists.

Also, one is not justified in saying, as Kuder (12) implies, that a person should score high on the mechanical scale in order seriously to consider engineering as a career. Chemical, civil, electrical, and sales engineers as groups do not have mean scores on the mechanical scale above the 65th percentile rank. Mechanical engineers and some industrial engineers did score significantly high on the average on the mechanical scale. The mean score of all professional engineers on this scale is below 65 P.R. (12, Table 2). Actually the interest typical of the large majority of engineers is characterized by significant elevations on scales 2 and 3 (Computational and Scientific).

Other discrepancies are apparent after com-

\* A revision of University of Missouri Counseling Bureau Research Report, No. 9a (Mimeographed), 1952. This report which included Tables 1 through 4 of the present paper is available from the authors at a cost of 50 cents per copy.



paring Kuder's groupings with empirical results. Thus, many of Kuder's original "logical" groupings now can be replaced by the increased body of empirical data. As a result a counselor can operate more effectively when he can base his test interpretation on real data.

In order to make this information based on actual test data usefully available to the counselor it is necessary to have it organized into some system. Wiener (20) proposed a coding system which coded each individual score over the 75th percentile rank. However, Frandsen (8) criticized Wiener's system as not being comprehensive enough; that is, the 75th percentile rank was too rigorous a cutting point. Instead of using only the scores above the 75th percentile rank, Frandsen suggested a coding system that would include deviations outside the 65th to 35th percentile rank range, and thus gain much better differentiation among the various occupations. By such a system, there would be less frequency of finding that the code for an individual's profile matches identically the codes of many different occupational groups. Also, a mean score which falls outside the 65th to 35th percentile range would be a significant deviation for almost any reasonably sized group.

Diamond (5) has shown how the use of a uniform cutting score for the Kuder scales is misleading and not in keeping with the reality of the occupational world as revealed in the census data. He notes that "not 25 per cent of employed urban men, but approximately 40 per cent, are engaged in occupations of a mechanical nature. It is, therefore, a statistical absurdity to expect that all men who enter the mechanical field shall have mechanical interest above the 75th percentile rank." On the other hand, there are some interest fields which employ only a fraction of one per cent of the labor force. Music is such a field. In this connection, Diamond points out that a musician who scores at the 75th percentile rank on the Kuder PR-V musical scale is more than two standard deviations below the mean of his occupational group.

So far, relatively little attention has been given to significantly low scores on interest test scales. The low scores may be equally useful in characterizing the interest of an occupational group as are high scores. For example, most engineers score significantly low on the social service scale.

It appears, therefore, that a system for coding Kuder PR-V profiles which would reflect the low as well as the high scores, would be quite helpful in studying the kind of interest which is typical of various occupational groups. Such a system should provide a code for a profile which is short and simple but which preserves a maximum amount of information.

A coding system is proposed here which meets Frandsen's (8) objections to Wiener's (20) system. It is similar to the system reported by Hathaway (10) and Holland *et al.* (11).

#### Coding Procedure

To code a profile, follow these steps. As an example we will use the percentile ranks corresponding to mean raw scores on the various scales of the Kuder PR-V made by a group of surgeons (12). The first number denotes the scale and the number after the dash denotes the percentile rank:

Out	Mec	Com	Sci	Per
0-68,	1-45,	2-25,	3-75,	4-27,
Art	Lit	Mus	Soc	Cle
5-61,	6-66,	7-62,	8-48,	9-25

Step 1. Select all scores of 75 P.R. and above and list the scale number in *descending* order of magnitude of the percentile ranks. Then place an apostrophe after these. Example: 3'.

Step 2. Select all scores between 74 P.R. and 65 P.R. inclusive and list the scale numbers in *descending* order of magnitude of the percentile ranks next after the apostrophe. Then place a dash after these. Example: 3'06—.

Step 3. Select all scores 25 P.R. and below and list the scale numbers in *ascending* order of magnitude of the percentile ranks. Then place an apostrophe after these. Example: 3'06—29'.

0-26.10 (D.O.T.)	Surgeon (job title)	52 (N)	3'06-29'4 (Code)
---------------------	------------------------	-----------	---------------------

Kuder PR-V, Form C, (P.R.):  
0-68, 1-45, 2-25, 3-75, 4-27, 5-61, 6-66, 7-62, 8-48, 9-25, V-

Reference: Kuder, F. G., Examiner Manual for the Kuder Preference Record  
—Vocational, Form C. Chicago, Ill.: Science Research Associates.  
February 1953. Table 2.

Notes: (description of the group, evaluation of the data, etc.)

FIG. 1. Example of card showing codes.

Step 4. Select all scores 26 P.R. to 35 P.R. inclusive and list the scale numbers in *ascending* order of magnitude of the percentile ranks next after the apostrophe. Example: 3'06—29'4.

Step 5. Place the V-Score in parentheses after the entries so far. This applies primarily to the coding of profiles of individuals. Example: 3'06—29'4 ( ).

It is proposed that any serious user of the Kuder PR-V prepare codes for all occupational groups for which profile data are available, such as those reported in the manual (12, Tables 2 and 3) and in various journal articles. Then a duplicate set of cards should be prepared for each occupational group. (See example.) One set of cards should be filed numerically according to code number and the other set alphabetically according to job title. Data for men and women should either be filed separately or on different colored cards.

*The Dictionary of Occupational Titles* (D.O.T.) code number is for cross reference to that system of classifying occupations.

Once these two files are prepared, any given profile can be coded and referred to the code file for a list of job titles which have similar codes. Also, the job title file can be searched to determine if the codes of the occupations being considered by the client agree reasonably well with the code of the client's profile. As new data become available, appropriate cards can be prepared and inserted in the card files.

Use of such a coding system facilitates

more valid use of the Kuder PR-V by bringing real data to bear on the interpretation of a profile rather than basing interpretation on "logical" guesses which have proved fallible in the past. It may be desirable to extend the code file to include individual cases so that the counselor may then refer to his own case records of individuals as well as occupational groups for aid in interpretation.

From the various sources of research, four tables of data have been compiled. Table 1 (M-code) lists the various *male* occupational groups according to the numerical value of their codes. The number of subjects in each group and the reference to the original data are also given. Table 2 (M-alphabet) lists the various *male* occupational groups alphabetically by job title. Table 3 (F-code) lists the various *female* occupational groups according to the numerical code value. Table 4 (F-alphabet) lists the various *female* occupational groups alphabetically by job titles.<sup>1</sup>

It must be remembered that knowledge that an individual has interest similar to a particular occupational group does not insure that he will be successful or even satisfied in that occupation. The power of Kuder PR-V to predict job success or satisfaction is largely

<sup>1</sup> Tables 1 through 4 have been deposited with the American Documentation Institute. Order Document No. 4322 from the ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress, Washington 25, D. C., remitting in advance \$1.75 for 35 mm. microfilm or \$2.50 for 6 × 8 in. photocopies. Make checks payable to Chief, Photoduplication Service, Library of Congress.

unknown. However, use of a system such as proposed here will bring us one step closer to prediction of job satisfaction and possibly to a lesser degree, job success.

There is a limitation in the use of codes of interest profiles when based on mean scores which should be borne in mind. If the interest of an occupational group is highly homogeneous, a single code will reflect this interest pattern quite accurately. However, if the interest of an occupational group is quite heterogeneous, a single code will not reflect the interest of that group accurately. Several codes, each based on a homogeneous subgroup, may be required to reflect accurately the interest of an occupational group.

An example of an occupational group which has heterogeneous interests is "secondary school teachers." The code for "all male secondary school teachers" and that of several of the sub-groups can be contrasted as follows:

all secondary school teachers (male)	'8—'14
commercial teachers (male)	9'2—1'3
mathematics teachers (male)	23'—4'9
social studies teachers (male)	8'6—1'5
music teachers (male)	7'6—123'
vocational training teachers (male)	'15—4'7

An example of an occupational group which appears to have homogeneous interests is "nurses." The codes for "all trained nurses" and several sub-groups are as follows:

all trained nurses (female)	8'3—'94
nurse educators (female)	8'3—9'4
general staff nurses (female)	'83—'94
private duty nurses (female)	8'3—'94
public health nurses (female)	8'—9'
supervisors and head nurses (female)	'8—'94

Researchers are urged to investigate and report on the "homogeneity of interest" when reporting on the interest of any occupational group. This may be accomplished by reporting the frequency of various codes which members of the group achieve. However, the

establishment of a code frequency distribution for an occupational group is often a difficult task. It can be done in several ways, the first of which might well be coding the mean scores. A second way might be a tabulation of how many persons in a group had scores on the various scales coded in different parts of the code; i.e., high (75 P.R. and above), near high (65–74 P.R.), low (25 P.R. and below), near low (26–35 P.R.), and not coded (36–64 P.R.). Table 5 is such a tabulation for 62 students in a nursing education program. It can be seen from Table 5 that 54 of the 62 student nurses scored 75 P.R. or above on scale 8 (social service) and none of them scored as low as the 35 P.R. Ninety-five per cent of this group scored 65 P.R. or above on scale 8. Thus we see that a reasonably high score on scale 8 is typical for almost all student nurses in this group. By similar analysis we can find other characteristics of our group, such as a low score on scale 9 (clerical).

The actual frequency with which any particular code occurs in a group is probably the most precise way in which to describe the interest of the group. However, this method does not lend itself well to the making of summary statements about a group. Of the 62 student nurses mentioned above, 18 of them achieved codes which contained an 83—94 code. That is, they may have had other scales coded high or low but scales 8 and 3 were coded high and scales 9 and 4 were coded low. Forty-one of the 62 student nurses had scales 8 and 3 coded high without regard for how other scales were coded. It required ten different codes or code variations to account for all cases in this group. However, eight of these ten codes were merely variations of the 83—94 code.

### Summary

A proposal has been made whereby interpretation of Kuder PR-V interest profiles can be made more valid by bringing to bear upon the interpretation the fast-growing body of empirical data relative to typical profiles for various occupational groups. A system for coding profiles has been described and some



Table 5

Frequency of Scores Appearing in the Various Parts of the Code for 62 Students in a Nursing Education Program

Kuder PR-V Scale		Position in the Code				
		High	Near High	Low	Near Low	Not Coded
0	Outdoor	32	7	10	4	9
1	Mechanical	15	6	15	6	20
2	Computational	12	1	26	8	15
3	Scientific	40	4	3	3	12
4	Persuasive	8	5	28	6	15
5	Artistic	6	9	18	10	19
6	Literary	10	4	19	13	16
7	Musical	17	7	12	6	20
8	Social Service	54	5	0	0	3
9	Clerical	2	1	41	7	10

ways of using codes in interpreting profiles have been presented. Finally, the effect of heterogeneity of interest within a group upon the use of a single code to describe the interest of that group was discussed as a limitation and a caution in the use of codes.

Received October 7, 1953.

### References

1. Baas, M. L. *A study of patterns among professional psychologists*. Unpublished M.A. thesis, Purdue University, 1949.
2. Baas, M. L. Kuder interest patterns of psychologists. *J. appl. Psychol.*, 1950, 34, 115-117.
3. Beamer, G. C., Edmonson, L. D., and Strother, G. B. Improving the selection of linotype trainees. *J. appl. Psychol.*, 1948, 32, 130-134.
4. Capwell, Dora J. *Psychological tests for retail store personnel*. Pittsburgh: Research Bureau for Retail Training, University of Pittsburgh, 1949.
5. Diamond, S. The interpretation of interest profiles. *J. appl. Psychol.*, 1948, 32, 512-520.
6. DiMichael, S. G. The professed and measured interests of vocational rehabilitation counselors. *Educ. psychol. Measmt*, 1949, 9, 59-72.
7. Eimicke, V. W. Kuder Preference Record norms for sales trainees. *Occupations*, 1949, 28, 5-10.
8. Frandsen, A. N. A note on Wiener's coding of Kuder Preference Record profiles. *Educ. psychol. Measmt*, 1952, 12, 137-139.
9. Hahn, M. E. and Williams, C. T. The measured interests of Marine Corps women reservists. *J. appl. Psychol.*, 1945, 29, 198-211.
10. Hathaway, S. R. A coding system for MMPI profiles. *J. consult. Psychol.*, 1947, 11, 334-337.
11. Holland, J. L., Krause, A. H., Nixon, M. E., and Trembath, M. F. The classification of occupations by means of Kuder interest profiles. *J. appl. Psychol.*, 1953, 37, 263-269.
12. Kuder, F. C. *Examiner Manual for the Kuder Preference Record—Vocational. Form C., Second Revision*. Chicago: Science Research Associates, 1951.
13. Lewis, J. A. Kuder Preference Record and MMPI scores for two occupational groups. *J. consult. Psychol.*, 1947, 11, 194-202.
14. Mathewson, R. H. and Herbert, R. *Kuder Preference Record Profiles for 48 occupational fields in six major groups*. Cambridge: The Guidance Center, 1949.
15. Rundquist, R. M. (personal communication). December, 1951.
16. Shaffer, R. H. The measured interests of business school seniors. *Occupations*, 1949, 27, 462-465.
17. Speer, G. S. The Kuder interest test patterns of fire protection engineers. *J. appl. Psychol.*, 1948, 32, 521-526.
18. Triggs, Frances O. Kuder Preference Record in the counseling of nurses. *Amer. J. Nursing*, 1946, 46, 312-316.
19. Triggs, Frances O. The measured interests of nurses. *J. educ. Res.*, 1947, 41, 25-34.
20. Wiener, D. W. Empirical occupational groupings of Kuder Preference Record profiles. *Educ. psychol. Measmt*, 1951, 11, 273-279.

## Transfer of Training in Tracking as a Function of Control Friction<sup>1</sup>

F. A. Muckler and W. G. Matheny

*University of Illinois*

The degree to which a training device should simulate the final task is of considerable practical interest to those concerned with training as well as to the manufacturers of training devices. A training device may simulate a psychomotor task to a greater or lesser degree along several dimensions. One of these dimensions is the control force necessary for accomplishing the task. The question becomes: what degree of fidelity of simulation of control force is necessary in the training device in order to secure optimum transfer of training?

The present study was designed to investigate the effect of varying control friction upon transfer of training in a visually guided tracking task.

Despite a considerable literature, the experimental evidence on the effect of friction in control mechanisms is not clear cut. In general, friction has been found to be undesirable, but the effect is a function of such variables as: (a) the type of friction involved (4, 7, 9, 13); (b) the tracking task (6, 8); (c) the presence or absence of inertia (7, 11); (d) the radii when handwheels or knobs are used (13, 14); and (e) the response measure recorded (6, 10, 15). Further, the effect of friction may be specific to complex interactions of many of these variables (7, 14).

All of these studies are concerned with either original learning or performance situations while the question of transfer from one control friction to a different control friction remains relatively uninvestigated. In a study summarized by Craik (2) and reported by Vince (16), subjects were trained to make

corrections with a lever operated against a stiff spring. After the subjects were making accurate movements, the spring tensions were changed. The new response was found to be delayed by at least 0.16 second; this time interval was termed the "kinesthetic reaction time." More directly applicable is the experiment reported by Bilodeau (1). Two groups rotated a crank handle at either heavy or light loads for five minutes. A third group practiced first under a light load and second under a heavy load alternately for one minute periods for five minutes. The fourth group started under a heavy load changing to a light load under the same procedure. Of interest here is the fact that when the latter two groups were shifted, "rate output was approximately equal to that of non-shifting groups" (1, p. 100). These data are interpreted here to imply that there is no specific effect of previous practice on either a heavy or light load to the performance of the task under the light or heavy load, respectively.

In this experiment, the effect of changing friction upon the level of performance in a visually guided tracking task was investigated. Experimental evidence was sought for a change from a higher friction to a lower friction, from a lower friction to a higher friction, and from a "frictionless" condition to a friction system.

### Experimental Method

*Experimental Task.* The task was following pursuit tracking and required the subject to track a continuous sine wave drawn along a moving roll of paper. This line passed behind a horizontal slit in a viewing panel at a rate of four cycles per minute. A lever-type control handle, moving horizontally forward and backward, controlled a pencil-type pointer. Tracking responses were recorded in the form of a continuous response line on the paper with the stimulus line.

The friction in the system could be varied systematically by means of a brake drum attached to the control lever. The friction was independ-

<sup>1</sup>This research was supported in part by the United States Air Force under Contract AF 33(038)-25726, monitored by the Air Force Personnel and Training Research Center. Permission is granted for reproduction, translation, publication, use and disposal in whole or in part by or for the United States Government. We should like to thank Dr. L. H. Lanier, Dr. A. C. Williams, and Dr. W. E. Kappauf for their valuable suggestions and criticisms.

ent of both rate and extent of control movement. Further, there was no "centering" tendency of the control lever.

*Procedure.* Each subject was given the following instructions:

This instrument is called a tracking device. In this opening (point) there will appear a moving line, which will go back and forth across the opening. This (point) is the control handle. As you move the handle forward, this pointer (show) will move to the right; as you move the handle backward, the pointer will move to the left. Now the pointer that you control will make a mark on the moving paper. Your job will be to match the mark you make with the line that is presented to you. Please use only one hand, the hand you start with. Are there any questions?

To reduce fatigue effects, the subjects were given a two-minute rest period after the completion of twenty cycles. After the subject had reached criterion on the original learning task, he was sent from the room while the control friction was changed. The time from completion of original learning to the beginning of the transfer trials was, in all cases, two minutes. To observe the effect of the two-minute break, the control group was given a two-minute rest and then continued the task with the same friction load.

*Criterion.* The subjects were said to have learned the pattern when they did not deviate more than two millimeters from the stimulus line for three successive cycles. A trial was defined as one sine wave cycle.

*Experimental design.* Seven experimental groups were assigned: 0 (approximately 2.5 ounces), 2, 4, 6, 8, 10, and 12 pounds. The basic design was the familiar paradigm cited by Woodworth (17) as Plan 4:

Transfer group learns A . . . . . Learns B  
Control group . . . . . Learns B

The control group selected was the six-pound friction group. Thus, three groups—8, 10, and 12 pounds—transferred to the lower control pressure of six pounds. The groups 0, 2, and 4 pounds transferred to the higher control pressure of six pounds. The basic design is the same in all cases.

Measurement of transfer is recorded in per cent savings of trials. The formula used is from Gagne, Foster, and Crowley (5):

Per cent transfer =  
$$\frac{\text{Control group score} - \text{transfer group score}}{\text{Control group score} - \text{total possible score}} \times 100.$$

Since the response measure used was the number of trials to criterion, the total possible score could be reduced to zero.

*Subjects.* A total of 105 Air Reserve Officer Training Corps Cadets were used. The age range was 17 to 25 years with a mean of 19.6 years. The subjects were assigned at random, on the basis of a table of random numbers (3), so that each experimental group contained 15 subjects. One restriction was placed on the randomization, namely, the subjects were assigned in blocks of seven.

## Results

*Original Learning.* The mean number of trials to reach criterion is shown in Table 1 for each experimental group. Since Bartlett's test for homogeneity (3) showed the variances of these scores to be homogeneous, and since the distribution of trials was found to be "moderately" normal, an analysis of variance was computed. There were no statistically significant differences between the experimental groups in original learning.

However, since the distributions did show some skewness, confirmation of the analysis of variance result was sought by the use of a distribution-free technique described by Mood (12) as "simple linear regression." The application of this test gave results completely in accord with those obtained from the analysis of variance.

*Transfer of Training.* The mean number of transfer trials necessary to reach criterion for every experimental group is shown in Table 1. Per cent transfer of training was computed on the basis of the formula mentioned previously. In Figure 1, per cent transfer of training is shown as a function of control friction. Individual transfer points are: 0 pounds, 86 per cent; 2 pounds, 91 per cent; 4 pounds, 90 per cent; 6 pounds (con-

Table 1  
Mean Number of Trials to Reach Criterion

Experimental Groups	Original Learning	Transfer Learning
0	28.6	3.8
2	25.5	2.5
4	27.0	2.7
6	27.3	0.0
8	25.6	1.4
10	22.0	2.5
12	24.9	2.8



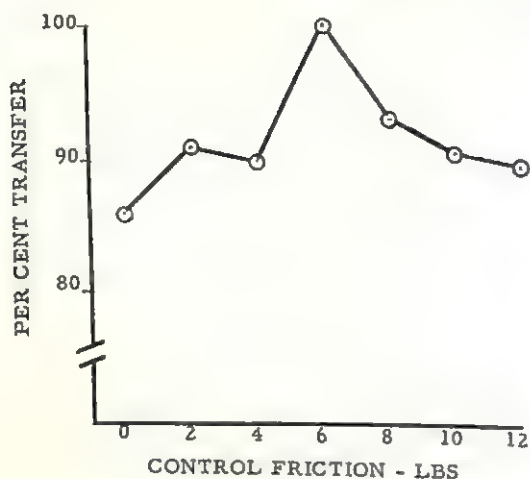


FIG. 1. Per cent transfer as a function of control friction.

trol), 100 per cent; 8 pounds, 93 per cent; 10 pounds, 90.8 per cent; and 12 pounds, 89.7 per cent.

It will be recalled that the control group (6 pounds friction) was given a two-minute break after original learning and then continued on the same task as may be seen in Table 1. There was no decrement of performance observed; the criterion level was maintained. This result may be interpreted as 100 per cent positive transfer and will be found, as such, in Figure 1.

Ignoring the 6 point control group, a test was made of the significance of differences between the experimental groups on raw score transfer scores. Since the distribution of transfer trials was highly skewed, the results were evaluated by the distribution-free technique previously described as "simple linear regression." The chi-square evaluation showed that the null hypothesis is accepted and that there were no statistically significant differences between the transfer groups.

### Discussion

**Original Learning.** The results indicate plainly that performance to criterion under these experimental conditions was independent of control friction with the response measure used. Of the literature previously cited, both Hick and Clarke (9) and Gray and Ellison (6) have obtained similar results.

**Transfer of Training.** The results indicate that a change in friction had very little effect on the level of performance. The lowest mean transfer for an experimental group was 86 per cent for the 0 pound group. Since there were no significant differences between experimental transfer groups, these data show that transfer of training in this tracking task was relatively independent of control friction.

The implication of these data for training devices seems clear. Where control forces are a variable, optimum transfer will be obtained by exact simulation; nevertheless, little will be lost if the control force varied. Obviously, since this conclusion rests on the results of a relatively simple laboratory task, further validation with specific training devices seems necessary.

### Summary

The effect of transfer from several amounts of friction to another level of friction in a manual control system was investigated. Transfer effect was found to range from 86 to 93 per cent positive transfer; it was found that transfer was relatively independent of control friction under the conditions used in this study. Finally, control friction had little apparent influence on original learning with the criterion measure used.

Received October 22, 1953.

### References

1. Bilodeau, E. A. Decrements and recovery from decrements in a simple work task with variation in force requirements at different stages of practice. *J. exp. Psychol.*, 1952, 44, 96-100.
2. Craik, K. J. W. Theory of the human operator in control systems. I. The operator as an engineering system. *Brit. J. Psychol.*, 1947, 38, 56-61.
3. Edwards, A. L. *Experimental design in psychological research*. New York: Rinehart, 1950.
4. Fitts, P. M. Engineering psychology and equipment design. In S. S. Stevens (Ed.), *Handbook of experimental psychology*. New York: Wiley, 1951, pp. 1287-1340.
5. Gagne, R. M., Foster, H., and Crowley, M. E. The measurement of transfer of training. *Psychol. Bull.*, 1948, 45, 97-130.
6. Gray, Florence E. and Ellison, D. G. Effects of friction and mode of operation upon accuracy of tracking with the GE pedestal sight. *AAF*,

- Air Mat. Comm., Aero Med. Lab. Report TSEAA-694-2c, 1947.
7. Helson, H. Design of equipment and optimal human operation. *Amer. J. Psychol.*, 1949, 42, 473-497.
  8. Hick, W. E. Friction in manual controls with special reference to its effects on accuracy of corrective movements in conditions simulating jolting. *Mot. Skills Res. Exch.*, 1949, 1, 9. (Abstract.)
  9. Hick, W. E. and Clarke, P. The effects of heavy loads on handwheel tracking. *Mot. Skills Res. Exch.*, 1949, 1, 20. (Abstract.)
  10. Jenkins, W. L., Mass, L. O., and Rigler, D. Influence of friction in making settings on a linear scale. *J. appl. Psychol.*, 1950, 34, 435-439.
  11. Jenkins, W. L., Mass, L. O., and Olson, M. W. Influence of inertia in making settings on a linear scale. *J. appl. Psychol.*, 1951, 35, 208-213.
  12. Mood, A. M. *Introduction to the theory of statistics*. New York: McGraw-Hill, 1950, 406-407.
  13. Raines, A. and Rosenbloom, J. H. Ideal torques for handwheels and knobs. *Machine Design*, 1946, 18(8), 145-148.
  14. Reed, J. D. Factors influencing rotary pursuit. *J. Psychol.*, 1949, 28, 65-92.
  15. Searle, L. V. and Taylor, F. V. Studies in tracking behavior. I. Rate and time characteristics in simple corrective movements. *J. exp. Psychol.*, 1948, 38, 615-631.
  16. Vince, Margaret. Corrective movements in a pursuit task. *Quart. J. exp. Psychol.*, 1948, 1, 85-103.
  17. Woodworth, R. S. *Experimental psychology*. New York: Henry Holt, 1938.

## A Correction of the Clark-Owens Validation Study of the Worthington Personal History Technique

Robert F. Peck

*Worthington Associates, Chicago, Illinois*

and

William Stephenson

*University of Chicago*

In a recent paper (1) on the Personal History method, the following conclusions were drawn:

1. Five isolated personality trait scores, from two standard inventories, were a more "efficacious" assessment device than individual reports derived by the Worthington Personal History method. "Efficacy" was not defined by the authors, but presumably they meant accuracy in predicting the job effectiveness and promotability of the industrial employees in the study.

2. This (they said) "constitutes damaging evidence as to the usefulness of the Personal History."

3. Furthermore (they continued) "these results . . . tend to follow the pattern of dubious or negative results found in validation studies of other projective techniques."

As a matter of fact, even the selected data included in the Clark-Owens report would lead an impartial investigator to exactly the opposite conclusion on each of these points.

Part of the explanation appears to lie in the fact that several major errors were committed in designing and executing this little study. Constructive corrections for these were recommended to Clark on January 6, 1953, following a conference with him in December, 1952; but the issues still appear to be disregarded in the recent Clark-Owens article.

1. The criterion was a set of ratings by co-workers (not supervisors), according to this pattern: Judges 1 and 2, in Dept. X, rated subject A; judges 3 and 4, in Dept. Y, rated subject B; and so on. No attempt was made to find the comparability of ratings made by different judges in different depart-

ments. Thus, the reliability of the criterion is an unknown quantity, with an error of unknown but undoubtedly considerable size. If it were not that these ratings proved significantly related to both the PH ratings and the standard inventories, implying some kind of meaningful stability in the criterion, this feature would invalidate the entire study.

2. The research was ultimately narrowed to a few traits, apparently because only these traits could be measured by the inventories. The proper procedure, of course, to validate the PH reports, would be to measure those traits which the PH covers. (Editorial policy does not allow space for an illustrative PH report. See reference 10, for an example.) The task of measuring the *interaction* of traits, which the PH undertakes to do, is beyond the scope of standard-inventory scores, of course, especially if the scores are taken singly. Perhaps for this reason, this latter issue was ignored. In short, the study was not really adequately designed to test the validity of the PH.

3. A peculiar and persistent error in using the Chi-square method is explained below in Conclusion No. 2.

4. Despite the fact that Clark and Owens (erroneously) termed the contingency coefficients for both PH and inventories "not significant," they proceeded to compare the coefficients for the two methods, though only on five personality traits. In doing this, they apparently did not realize that contingency coefficients from different sets of data cannot be compared unless a class-index correction is applied (2). These are not correlation coefficients. Without the correction, it is impossible to tell whether a C of .75 from one



set of data is larger, equal, or smaller than a C of .65, .75 or .85 from different data. This is a relatively minor point, but it is still an error.

#### The Correct Conclusions from the Data

Despite the questionable or fallacious procedures, the actual data which Clark and Owens report clearly show the following facts:

1. The Personal History reports were translated into personality-trait ratings and into job performance ratings, by five psychologists, with a high degree of reliability (Adjustment to Others .91, Job Effectiveness .79, Promotability .93, for example).

2. The Personal History ratings thus obtained showed a high, significant relationship to the criterion, both on the personality traits and on Job Effectiveness, Adjustment to Co-workers, and Promotability.

#### Contingency Coefficients (C) PH vs. Ratings

Active	.605
Impulsive	.655
Dominant	.654
Stable	.676
Sociable	.585
Job-effectiveness	.513
Promotion Possibilities	.697
Adjustment to Others	.614

Through a misuse of chi-square methods (pointed out to them in the letter of January 6, 1953), the authors report that these contingency coefficients, ranging from .51 to .70, are "not statistically significant." Mr. Clark reported, in December 1952, that this happened because the 47 subjects were subdivided into many cells, several of which contained less than 5 cases. Since an extremely large correction factor has to be applied—a procedure which is not acceptable, even technically, to most statisticians—almost *no* coefficient would appear significant, no matter how high. This is a technically possible, but logically meaningless, procedure. However, their own findings indicate that if proper chi-square divisions were applied to these data, both the Personal History and the standard tests would show a significant degree of relationship with the criterion ratings. This, at least, is our considered opinion, and that of

several other statistically competent psychologists (3, 4).

3. The standard inventories showed a significant relationship to the criterion on *five isolated personality traits*, of about the same order as the PH-criterion relationship on these five traits (Active, Impulsive, Dominant, Stable, Sociable). However, *these inventory trait scores show no power to predict Job Effectiveness, Adjustment to Co-workers, or Promotability*. Indeed, it appears from the Clark-Owens report that no effort was made to attempt such a prediction from the inventories, although the criterion was available.

4. Thus, on the crucial criteria for determining the efficacy, as well as the validity, of any assessment method (5)—Adjustment to Co-workers, Job Effectiveness, and Promotability—the Clark-Owens data show that the Personal History method was significantly effective. Since the authors report no attempt to measure the predictive power of the standard inventories against these criteria, the "efficacy" of those inventories for predicting job performance remains wholly untested and unproven. Indeed, since the PH measured the individual traits about as well as the inventories, and additionally measured job performance, it would seem that the inventories are not needed, in this setting. This is contradictory, of course, to the statements Clark and Owens made about "efficacy."

5. Clark and Owens' remark about "dubious or negative" findings on other validation studies of projective techniques requires reference to numerous studies which have demonstrated positive validity for these methods. Naïveté, or errors of logic, in research design and in the use of statistical methods, have frequently resulted in "dubious" findings. However, properly designed research has repeatedly shown that projective techniques, among them the Personal History method, can be valid predictors of overt, daily behavior in the work world, as well as in the clinic (6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16).

Received June 4, 1954.

Published out-of-turn by the editor.

## References

1. Clark, J. G. and Owens, W. A. A validation study of the Worthington Personal History blank. *J. appl. Psychol.*, 1954, 38, 85-88.
2. Kelley, T. L. *Statistical method*. Macmillan Co., N. Y., 1924, p. 266.
3. Cureton, E. E. Validity, reliability and baloney. *Educ. psychol. Measmt.*, Spring, 1950, 10, 94-96.
4. Lewis, D. and Burke, C. Use and misuse of chi square technique. *Psychol. Bull.*, 1949, 46, 433-489.
5. Thorndike, R. L. *Personnel selection*. John Wiley & Sons, N. Y., 1949.
6. Beck, S. J. *The six schizophrenias: reaction patterns in children and adults*. Res. Monog. No. 6, Amer. Orthopsychiatric Assoc., Inc., N. Y., 1954.
7. Endacott, J. I. *Methodology for the study of clinical cases by the way of Rorschach and psychoanalytic theories*. Unpublished Ph.D. dissertation, The University of Chicago, 1954.
8. Henry, W. E. *The Thematic Apperception technique in the study of culture-personality relations*. Genet. Psychol. Monog., 1947, 35, 1-135.
9. Nevis, E. C. *The effectiveness of the Worthington Personal History technique in assessing Air Force officers for command and staff leadership*. Unpublished Ph.D. dissertation, Western Reserve University, 1954.
10. Peck, R. F. and Thompson, J. M. The use of individual assessments in a management development program. *J. personnel admin. industr. Relat.*, April, 1954, 1, 79-98.
11. Peck, R. F. and Worthington, R. E. New technique for personnel assessment. *J. personnel admin. industr. Relat.*, January, 1954, 1, 23-30.
12. Spencer, G. J. and Worthington, R. E. Validity of a projective technique in predicting sales effectiveness. *Personnel Psychol.*, 1952, 5, 125-144.
13. Stephenson, W. Q-methodology and the projective techniques. *J. clin. Psychol.*, 1952, 8, 219-229.
14. Swint, E. R. The Worthington Personal History: a report. *J. ind. Train.*, Nov.-Dec., 1950.
15. Swint, E. R. and Newton, R. A. The Personal History—a second report. *J. ind. Train.*, Jan.-Feb., 1952.
16. Worthington, R. E. Use of the Personal History form as a clinical instrument. Unpublished Ph.D. dissertation, The University of Chicago, June, 1951.

## A Reply to Drs. Peck-Stephenson

William A. Owens, Jr.

Iowa State College

Drs. Peck and Stephenson, as might have been anticipated from their obvious interest, have seen fit to make some interesting and ingenious comments upon the Clark-Owens study (1) of the Worthington Personal History Blank (PH). However, since their comments purport to be "a correction" they should be examined in order.

1. Peck and Stephenson say the criterion employed, that of associates' ratings, is unreliable, although they state that it "proved significantly related to both the PH ratings and the standard inventories." Actually, these relationships were *not* statistically significant, although this "unreliable criterion" was found to be more closely related to standard inventory results than to PH results (on the only five traits presumably measured by both) five times out of five. In this regard, at least, it was quite consistent.

2. Peck and Stephenson seem to feel that the five traits common to PH and the available standard inventories were not enough to constitute any real evidence as to the validity of PH. It was, of course, only in the case of these five traits that PH could really be evaluated, since low criterion relationships could well be attributed to low criterion reliability or validity unless they were *differentially* low. They also state that the PH measures the interaction of traits (we presume clinically, since no quantitative evidence is quoted), and that it, therefore, goes beyond the scope of standard inventories in a global direction. However, the obtained Clark-Owens estimate of the relationship between PH results and criterion ratings is lowest in the case of "job effectiveness"—a complex characteristic—and about average for "promotion possibilities" and "adjustment to others." It would thus seem that the PH does *not* yield better global than simple estimates, in this sample.

3. Peck and Stephenson accuse Clark and Owens of a "peculiar and persistent error in

using the Chi-square method," in spite of their earlier advice to us. Let us examine their arguments. (a) They say that we divided our 47 cases among too many cells, "several of which contained less than 5 cases." However, the theoretical consideration relates to *expected* frequencies, *not* to observed, and even so, the number 5 is relatively arbitrary (5). (b) They imply that some enormous correction for continuity should have been made and was ignored, whereas Cochran (2) states that "Tables with more than 1 degree of freedom and some expectations greater than 5—should—use  $\chi^2$  *without correction for continuity*." (c) Finally, they conclude that, in their considered opinion, both the PH and standard inventory results would be significantly related to the criterion in, say, a  $2 \times 2$  table. How this could happen is a bit hard to understand, since Guilford (3) states, "There is probably nothing to be gained by applying Yates's correction when there is more than 1 degree of freedom." And again, still quoting Guilford, "The effect of the correction is to *reduce* the size of  $\chi^2$ ." Thus, *had Clark-Owens followed the procedure suggested by our critics, the effect would have been to remove the obtained  $\chi^2$  values still further from significance.*

4. Peck and Stephenson say, quite correctly, that contingency coefficients cannot be compared without making a class-index correction. They also say, even more correctly, that "this is a relatively minor point." Actually, making this correction would do practically nothing to the relative magnitudes of PH vs. test validities. The test coefficients would tend to receive larger corrections, since they are initially larger; and the PH would tend to receive larger corrections because the number of cells is somewhat smaller. If Peck and Stephenson had bothered to compute it, they could have observed that the differential shifts could not have exceeded .01 or .02. This, of course, would not remotely approach



changing the direction of a single difference—and direction is all that is involved in the randomization test.

5. In their second section, purportedly dealing with "Correct Conclusions from the Data" Peck and Stephenson become very seriously, if not willfully, confused. They appear to mistake an omission for a negative result saying, "these inventory trait scores show no power to predict Job-Effectiveness, Adjustment to Co-workers, or Promotability." The reason they do not is that, in an attempt to be fair to the PH method, Clark-Owens did not report them. Actually, three of our judges subsequently did considerably better in predicting these three characteristics from the objective test results than from the PH. However, they were more familiar with the former, and the data may have been slightly contaminated. In any case, it was Clark-Owens' stated purpose to evaluate PH vs. objective tests—not PH vs. objective tests *plus* an imponderable interpreter of them. Peck and Stephenson surely realize that the tests do not yield scores on these three characteristics.

6. Finally, Clark-Owens' critics take them to task for a comment about "the pattern of dubious or negative results found in validation studies of other projective techniques." An answer to them requires only a reference to Schofield (4), who summarizes all validity studies reported in 1949, 1950, and 1951, and indicates that two-thirds to three-fourths of them yielded negative results.

All-in-all, Clark-Owens must firmly reject the alleged corrections of Peck-Stephenson, although fully granting the limitations of their study as originally set forth.

Received July 20, 1954.

Published out-of-turn by the editor.

#### References

1. Clark, J. G. and Owens, W. A. A validation study of the Worthington Personal History Blank. *J. appl. Psychol.*, 1954, 38, 85-88.
2. Cochran, W. G. The  $\chi^2$  test of goodness of fit. *Ann. Math. Statist.*, 1952, 23, 315-345.
3. Guilford, J. P. *Fundamental statistics in psychology and education*. New York: McGraw-Hill, 1950.
4. Schofield, W. Research in clinical psychology. *J. clin. Psychol.*, 1952, 8, 255-261.
5. Walker, Helen M. and Lev, J. *Statistical inference*. New York: Henry Holt, 1953.

## Applied Psychology in Action

### GATB in Foreign Countries

Beatrice J. Dvorak

*Testing Branch, U. S. Employment Service, Washington 25, D. C.*

The USES General Aptitude Test Battery has been translated into a number of foreign languages, and research is being conducted in these foreign countries to adapt and standardize it for use on populations in those countries. Permission has been granted by the U. S. Employment Service to the following organizations and individuals to use the GATB in such research. While information is not available regarding the status of all of these projects, it is known that the French, Japanese, Portuguese, and Spanish editions have already been published.

#### Argentina

Carlos A. Pourteau Agote  
Universidad de Buenos Aires  
Laboratorio Psicotecnico  
Republica, Argentina

#### Australia

H. A. Bland  
Department of Labour and National Service  
Melbourne, Australia

#### Belgium

R. Buyse  
University of Lorraine  
Tournai, Belgium

Jean Herickx  
Centre d'Orientation  
Bruxelles, Belgium

M. Dewals  
Psychotechnicien de la Société Nationale  
des Chemins de Fer Vicinaux  
Bruxelles, Belgium

Capitaine Commandant Hourman  
Chief du Centre d'Orientation  
Ministère de la Defense Nationale  
Bruxelles, Belgium

F. Vandenborre  
Ministère de l'Instruction Publique  
Bruxelles, Belgium

#### Brazil

Jacy Magalhaes  
Divisao de Organizacao do Trabalho  
Rio de Janeiro, Brazil

Livraria Oscar Nicolai  
Caixa Postal 246  
Brazil

S. J. Schwarzstein  
Director do Servico de Colocacao e Informacao  
Profissional  
Sao Paulo, Brazil

Secretaria do Trabalho  
Servico de Colocacao e  
Informacao Profissional  
Sao Paulo, Brazil

#### Canada

Morgan D. Parmenter  
Director, The Guidance Centre  
University of Toronto  
Toronto 5, Canada

#### China

Ministry of Social Affairs  
Shanghai, China

#### Denmark

Poul Bahnsen  
Director, Psykotekniske Institut  
Copenhagen K.—Denmark

Paul Vidriksen  
Arbejdsdi Rektoratet  
Kopenhagen, Denmark

#### England

M. Desai  
Psychological Department, London County  
Council  
London, England

H. J. Eysenck and J. Tizard  
The Maudsley Hospital  
London S. E. 5, England

C. B. Frisby  
Director, National Institute of Industrial Psy-  
chology  
London W. C. 2, England

Roland Harper and D. R. Martin  
The University of Leeds  
Leeds 2, England

B. W. Richards  
St. Laurence's Hospital  
Caterham, Surrey, England

Constance M. Mathieson  
East Anglian Regional Hospital Board  
Norwich, Norfolk, England

Alec Rodger  
Birkbeck College, University of London  
London, England

#### India

Vocational Guidance Bureau  
Bombay, India

#### Italy

Ing. Vincenzo Flagiello  
Societa per l'Industria e l'Elettricit   
Centro Istruzione Professionale  
Viale Benedetto Brin  
Terni, Italy

Agostino Gemelli  
Director, Laboratorio di Psicologia Sperimentale  
Milano, Italy

Guido Majaron  
Viale Arnaldo Fusinato 2F  
Vicenza, Italy

Consiglio Nazionale delle Ricerche  
Istituto Nazionale di Psicologia  
Rome, Italy

#### New Zealand

Auckland University College  
Auckland C. 1, New Zealand

W. J. H. Clark  
Vocational Guidance Centre  
Auckland, New Zealand

#### Peru

Santiago Salinas  
Ministerio de Trabajo y Asuntos Indigenas  
Lima, Peru

#### Philippines

Antonio V. Roxas  
Escolta, Manila, Philippines

#### Scotland

P. S. Boyd and W. M. Miller  
Department of Mental Health  
Aberdeen, Scotland

#### South Africa

D. J. Du Plessis  
Department of Labor  
Johannesburg, Union of South Africa

C. P. J. Erasmus  
University of the Orange Free State  
Bloemfontein, South Africa

Department of Psychology  
University of Stellenbosch  
Stellenbosch, South Africa

Evryl Fisher  
Church Street  
Cape Town, South Africa

#### Sweden

Torsten Husen  
Cintrala Varnpliktsbyran  
Personalprovningsdetaljen  
Stockholm 10, Sweden

#### Switzerland

J. F. Herzog  
Office d'Orientation Professionnelle  
Neuch tel, Switzerland

Ph. H. Muller  
Universit  de Neuch tel  
Neuch tel, Switzerland

#### Turkey

Faruk Kardam  
Director-General of the Turkish Employment  
Service  
Ankara, Turkey



## Book Reviews

Tuckman, J. and Lorge, I. *Retirement and the industrial worker: prospect and reality*. New York: Bureau of Publications, Teachers College, Columbia University, 1953. Pp. xvi + 105. \$2.75.

This book reports the results of a survey undertaken at the request of the New York Cloak Joint Board of the International Ladies' Garment Workers' Union. The study investigated, by means of personal interviews, the attitudes toward retirement of three different groups of persons. These groups consisted of (1) 204 men and women still on their jobs, (2) 216 men and women who had submitted applications for retirement but who were still working, and (3) 240 retired persons. All interviewees were or had been members of the above named union, and all had earned their livelihood in the needle trades. The schedules used in the interviews were designed to obtain information relative to a wide variety of employment-retirement questions which fall generally under six main headings. These headings form the outline for the book and include: Retirement Attitudes, Health, Pressure Effect of Aging on Work Performance, The Worker's Preparation for Retirement, Effect of Retirement on the Family, and Factors Related to Retirement Attitudes.

Results are, of course, reported in terms of percentages of respondents falling in each of various response categories. Statistical significance is tested by means of chi square tests. It is to the authors' credit that they stay close to their facts and figures. They do not commit the error (so common in the literature about the problems of older employees) of launching into long opinionated discussions. Nor do they attempt to derive generalizations from their data which are not warranted by the narrowness of the population studied.

The last pages of the book consist of an excellent summary of the study and a short section devoted to conclusions and recommendations. It is this last section that will prove most useful to other persons doing research on older employee utilization. For it is here that one finds a wealth of hypotheses that need to be tested on a broader basis.

The primary barriers to the utilization and/or happy retirement of older persons are clearly outlined and questions are formulated which could well form the framework for other research programs designed to find methods of overcoming these barriers.

Presentation of the survey results could have been made much clearer and more easily understood. As it is, the reader is confronted with table after table of percentages which, although clearly titled and well organized, finally contribute to an overwhelming sense of boredom. Simple bar diagrams, pie charts, frequency polygons, and histograms could have been used to great advantage to facilitate quick and accurate interpretation of the results presented.

This book represents one of the most extensive researches into the attitudes and problems of working, retiring, and retired workers yet performed. As such, it is a "must" for persons engaged in the study of employment and retirement problems of older employees. In addition to the wealth of data presented, it is a rich source of research hypotheses, and points up the problems which must still be solved by researchers in this area.

Marvin D. Dunnette

*Industrial Relations Center  
University of Minnesota*

Berdie, R. F. (Editor). *Roles and relationships in counseling*. Minnesota Studies in Student Personnel Work, No. 3. Minneapolis: University of Minnesota Press, 1953. Pp. 37. \$1.25.

This publication consists of three papers presented at the Second Annual Conference of Administrators of College and University Counseling Programs held at the University of Illinois in 1951.

In the first paper, John Gustad discusses the definition of counseling. Clinical psychology and counseling can be considered to be essentially "one general kind of endeavor but with differing emphasis." Both include psychotherapy "where appropriate to the client and within the province of the practitioner." Teaching and counseling are differentiated largely in terms of different training and experience. His definition of counseling stresses the role of learning and the

requirement of professional competence. The analysis of the problem and review of the literature are helpful.

In the next paper Ralph Berdie describes the tactics and techniques developed in the Counseling Bureau at Minnesota to deal with problems of human and institutional relationships (which he terms public relations "in a very limited sense"). His base point is effective service. Beyond this he describes a number of gambits to improve client relationships and outlines an equally active program to promote intra- and extra-institutional staff relationships. Counseling administrators will find a number of suggestions for performance of a function rarely discussed in print. Unfriendly voices will perhaps find evidence to reinforce their suspicions of "empire-building."

Harold Pepinsky's concluding paper argues the thesis that the counseling psychologist "can help to build a culture in which the individual members are able to communicate with each other, to respond positively to each other, and to work together toward common group objectives," and provides a rationale for the use of group procedures in pursuit of this aim. Such activities appear to be geared to reaching a much larger proportion of the student body in the interests of community-wide mental health.

On a limited sector, the college or university campus, these papers exemplify a phenomenon of our times—the development (including growing pains) of new professional groups eager to provide experiences aimed at helping man to cope with the "human predicament." It is understandable that counseling psychologists are zealous and ambitious; the need is great.

Arthur H. Brayfield

Kansas State College

Bross, Irwin D. J. *Design for decision*. New York: Macmillan, 1953. Pp. viii + 276. \$4.25.

Every so often a book arrives for me to review that I find interesting and exciting. Bross's *Design for decision* is one of the few that falls into this category. It was read hastily once with enthusiasm and unflagging interest and almost without interruption of any sort. I could hardly wait to finish one

chapter so that I could move on to the next. During the weeks that followed the first reading, I picked up the book many times to read various sections at a more leisurely pace and my enthusiasm for it has not diminished to any noticeable extent.

It may be granted that the book introduces nothing that is not the common knowledge of all modern statisticians. But it does say what it has to say in a manner that few, if any, other modern statisticians have been able to say it in. Bross can write and he writes very well indeed.

Do not let that word "statisticians" that I used above mislead you. This is not a book about statistics in the sense in which you may interpret that word. It is not, for example, a collection of formulas, illustrative calculations, mathematical derivations and proofs. As the author states, no mathematics is required for reading it other than high school algebra. As a matter of fact, even if you have forgotten your high school algebra, you'll still get along with the text pretty well. Rather, this is a book about decision making or, more precisely, about statistical decision.

What is statistical decision? You may get some indication of what statistical decision involves from the following listing of chapter headings: history of decision, nature of decision, prediction, probability, values, rules for action, operating a decision-maker, sequential decision, data, models, sampling, measurement, statistical inference, statistical techniques, design for decision.

My answer, although admittedly inadequate, is that statistical decision is a method for making decisions that has its origins in a variety of specialized fields. I might even go so far as to identify statistical decision with scientific method, though Bross may not agree with this viewpoint. Anyway, the best answer as to what statistical decision is can be obtained by reading Bross's book for yourself.

I should add that there is something in this book for everyone. If you have no statistical training, *Design for decision* will tell you how a modern decision-maker operates—without overwhelming you with mathematical details. If you have some experience in applied statistics, then, as Bross points out, you may find that this book "provides a vantage point from which it is possible to see all of the



scattered techniques in their proper perspective."

"Some readers may be intrigued by the ideas of Statistical Decision because they represent a new advance toward the solution of a basic human problem. The principles have a wide scope; they apply to the choice of a foreign policy or to the private decisions that we all must make. They are, if you like, philosophical principles, a way of looking at the world in which we live, a guide to action in that world."

If the above paragraph from the introductory chapter of Bross's book doesn't whet your appetite and stir you to march out to your nearest library or bookseller for a copy of *Design for decision*, then you are a lost cause and no additional words of praise of mine for this book are going to help.

Allen L. Edwards

*The University of Washington*

Remmers, H. H. *Introduction to opinion and attitude measurement*. New York: Harper & Bros., 1954. Pp. 437. \$5.00.

Prepared as a college textbook, this volume by Dr. H. H. Remmers, professor of psychology and director of the Division of Educational Reference at Purdue University, offers a panoramic view of the field of opinion and attitude measurement, with emphasis divided between method and application.

"The realization is rapidly growing," the author says, "that attitudes, the way individuals and groups feel about the various aspects of their world, are probably more determinative of behavior than mere cognitive understanding of this world."

Part I is devoted to a discussion of techniques, including sampling and statistical theory, scaling, single question evaluation, the "summated questionnaire," and some of the "less direct measures of attitudes"—projective methods, sociometric approaches, rating scales, and the concept of empathy.

In Part II, Dr. Remmers describes many of the varied uses to which attitude and opinion measurements have been put in business, industry, the government, the study of community interrelations, and education.

The book is fairly comprehensive, succinct, and—in the main—a readable presentation. Appended to each of its dozen chapters are a

brief critical summary, a list of questions, and a bibliography.

The chapter on scaling techniques contains an able exposition of the Thurstone and Likert contributions, moves on to scale analysis as developed by Guttman, and deals in some detail with "the Cornell technique," because, says Dr. Remmers, it appears to be the one "most likely to be feasible in the greatest variety of situations."

His description of personality, interest, and problem inventories is of equal merit. He offers a quite lengthy report on the procedures that were followed in developing the Science Research Associates' Youth Inventory, under auspices of the Purdue Opinion Panel, with which the author is identified.

Dr. Remmers treats extensively of the applications of attitude and opinion measurements by educators. He reviews also the utilization of similar methods by the businessman to improve his advertising programs and his products; by industry in the study of employee attitudes, plant morale, absenteeism, and workers' opinions of members of minority groups; by social researchers in the analysis of intergroup and interpersonal relationships.

Unhappily, the book appears to lack freshness. Some of the material is obviously "dated"; one gains the impression that the author, except in a few instances, stopped collecting data along about 1947 or 1948, though much that is worth while has appeared in the literature since then. To illustrate Census Bureau sampling, he describes what was done in the 1940 census; to describe the government's uses of attitude and opinion studies, he dwells on World War II operations; to indicate the scope and nature of the Survey of Consumer Finances, he discusses what was done in the first year, 1946. With a mild apology for its absence, the author omits any material on how the television industry has put social science research to work.

The implication conveyed by the word "Introduction" in the title, that this is a textbook for beginners, may be somewhat misleading; it quickly becomes apparent that the college student will find himself in deep water unless he has been forearmed with preliminary work in statistics and psychology.

Notwithstanding, the volume is a scholarly



and well-planned treatise. In writing it, Dr. Remmers has made a substantial contribution toward effecting the kind of "popular understanding of the importance and implications" of the findings of the social scientists which he, at the outset, urges. The book is one of Harper's Psychological Series, of which Gardner Murphy is editor.

Sidney S. Goldish

*Minneapolis Star and Tribune*

Sherif, Muzafer and Wilson, M. O. (eds.). *Group relations at the crossroads*. New York: Harper, 1953. Pp. viii + 379. \$4.00.

Like the preceding *Social Psychology at the Crossroads*, this volume is a collection of papers emphasizing social-psychological concepts and explanations, prepared for a conference at the University of Oklahoma (April, 1952).

Sherif begins with an excellent summary introduction. Next comes J. P. Scott's "Implications of Infrahuman Social Behavior for Problems of Human Relations." This is one of the few fairly extensive reviews of the literature in the book. Scott uses the concept of levels of organization to describe phylogenetic differences in social behavior.

In considering "Psychological Traits and Group Relations," Anne Anastasi traces in detail the changes in approach in the area from a quest for a racial hierarchy to a more sophisticated multiple trait approach emphasizing the use of analysis of variance and factor analysis in which interactions between various traits and groups are expected to exist.

Anselm Strauss' "Concepts, Communication and Groups" discusses the primary importance of language in the development of human social behavior.

J. J. Gibson's "Social Perception and Psychology of Perceptual Learning" is an outline of the process of perceptual learning in terms of generalization and differentiation.

Gardner Murphy's "Knowns and Unknowns in the Dynamics of Social Perception" considers the importance of differential group membership on differential perception, the lines of cleavage between groups, and their significance.

"Development of the Small-Group Research Movement" by R. E. L. Faris covers mainly sociological studies in the area.

Herbert Blumer's "Psychological Impact of the Human Group" is a reiteration of the need to study both the group situation and the individual in developing an adequate theory of human interaction. An attempt is made to spell out some significant elements of group behavior.

Sherif's paper on reference groups reiterates the importance of reference groups as distinct from membership groups in understanding social behavior.

Leon Festinger's "An Analysis of Compliant Behavior" is a discussion of the empirical validity of two hypotheses: (1) public compliance without private acceptance will occur if the person in question remains compliant and in the group to avoid punishment; (2) public compliance with private acceptance will occur where it is satisfying to remain with those influencing the person.

Launor Carter's "Leadership and Small Group Behavior" is primarily a summary of his experimental studies on the behavior of leaders, the generality of leadership, and the effects of the group on leader behavior.

The need to consider "social distance" relatively, and to differentiate it from prejudice, is the theme of Mozell Hill's paper.

Nelson Foote and Clyde Hart's "Public Opinion and Culture Behavior" point to: (1) the dangers of depending only on poll answers to gauge public opinion; (2) the possibilities of analyzing public behavior in order to assess public opinion.

Helen Jennings concludes with a survey combining conclusions from sociometric studies with, as yet, unpublished sociodramatic examples to point out their significance for understanding personality and group formation.

Despite the heterogeneity of aims and methods of each of the papers, the reviewer will hazard presenting some overall impressions of the book.

1. While such general psychology topics as perceptual development have been included, contributions from many of the largest research programs on group relations such as the Ohio State Leadership Studies and the Michigan Survey Research Center are excluded. Anthropology is also absent.

2. Some of the papers could have been more of a contribution had they more extensively surveyed the literature. Yet, in gen-

eral, most papers tended to maintain a high standard of excellence.

3. Social psychologists seem to be trying hard to adopt sociological concepts and to integrate their work via "interdisciplinary" research with the other social sciences. It may be both more parsimonious and profitable for social psychologists to integrate their research with the general psychology of learning, perception and motivation. This does not mean that rejecting many sociological concepts will lead to ignoring the nature of the stimulating situation while studying social behavior. Rather, the situation will be described in terms which will lend themselves more readily to integration with psychological concepts describing the other equally important determinants of social, and all behavior, i.e., the behavioral history, motivation and biological level of the behaving organisms.

Bernard M. Bass

*Louisiana State University*

Schlotter, Bertha E. and Svendsen, Margaret. *An experiment in recreation with the mentally retarded.* (Rev. ed.) State of Illinois, Department of Public Welfare. Published by National Mental Health Funds. 1951. Pp. 142. Gratis.

This book is a re-issue of the volume published in 1932. Additions are a new introduction by the director of the Department of Public Welfare, and a thirteen page preface by Bertha E. Schlotter which provides an overview of the continuing effects of the recreational program begun in 1929. Requests from other institutions for copies of the earlier publication led to this re-issue.

The Illinois Institute for Juvenile Research, in a survey of the recreational program at the Lincoln State School and Colony in 1929, reported institutional overcrowding, inadequate facilities and staff, poor use of facilities, and overemphasis on maintaining quiet and order. Recreational activities provided for active participation by only 100 of the 2,600 patients then under care. The establishment of a department of recreation and a one-year experiment with a recreation program on an institution-wide basis followed. This book discusses staff qualifications, in-

service training programs, grouping of patients, and equipment and space problems. Specific lists of equipment, musical selections, books, and activities are included, with comments concerning their use and modification. About half of the book is devoted to "socio-psychological analysis of play activities." This section classifies activities in several ways: alphabetically, with minimum MA indicated; grouped for several MA levels; according to the degree of motor activity; according to need for equipment; and according to type of social organization and participation.

There are some important values of the book: the inclusion of lists of source books for games, songs, activities, and dances is of special interest to the recreational worker; the beginning worker will benefit by the vicarious experience made available. There are useful suggestions for modifying activities to suit special needs, and "leads" as to the handling of difficult patients. There is a real exemplification of the wide practical implications of the concept of individual differences in work with defectives.

From the viewpoint of this reviewer, it is unfortunate that the author attempted in the preface to defend the program in terms of psychological "principles" which are often inconsistent and contradictory, and sometimes not principles at all. The psychologically unsophisticated reader might be over-impressed by the comment, "This belies the belief that punishment, drill, and rewards are justified in the treatment of mental defectives" (p. 12). Comments concerning the level of performance attained would also be misleading to the neophyte in the field of mental deficiency: i.e., "In their dancing they show skill, variety, imagination, and spontaneity" (p. 17). Without at least a reminder to the reader of relative standards of expectation, such statements are potentially dangerous.

A recreation worker interested in the mentally deficient should study this book, but should maintain a cautious attitude toward the generalizations while embracing the specific helpful suggestions and making full use of the source material.

Harriet E. Blodgett

*Institute of Child Welfare,  
University of Minnesota*



## New Books, Monographs, and Pamphlets

Books, monographs, and pamphlets for listing and possible review should be sent to Dr. John G. Darley, Editor-elect, Graduate School, University of Minnesota, Minneapolis 14, Minnesota.

- Intelligence.* L. J. Bischof. New York: Doubleday and Company, Inc., 1954. Pp. 33. \$.85.
- Fundamentals of psychoanalytic technique.* Trygve Braatoy. New York: John Wiley & Sons, Inc., 1954. Pp. 404. \$6.00.
- Introduction to psychiatry.* O. Spurgeon English and Stuart M. Finch. New York: W. W. Norton & Company, Inc., 1954. Pp. 621. \$7.00.
- Nature and nurture: A modern synthesis.* John L. Fuller. New York: Doubleday and Company, Inc., 1954. Pp. 40. \$.85.
- Methods of research.* Carter V. Good and Douglas E. Scates. New York: Appleton-Century-Crofts, Inc., 1954. Pp. 920. \$6.00.
- Community and environment.* E. A. Gutkind. New York: Philosophical Library, 1954. Pp. 81. \$3.75.
- Social planning in America.* Joseph S. Himes. New York: Doubleday and Company, Inc., 1954. Pp. 59. \$.95.
- The deaf and their problems.* Kenneth W. Hodgson. New York: Philosophical Library, 1954. Pp. 364. \$6.00.
- Guidance services.* J. Anthony Humphreys and Arthur E. Traxler. Chicago: Science Research Associates, Inc., 1954. Pp. 438.
- Perception.* William H. Ittelson and Hadley Cantril. New York: Doubleday and Company, Inc., 1954. Pp. 33. \$.85.
- Social psychology.* Revised Edition. Otto Klineberg. New York: Henry Holt and Company, 1954. Pp. 578. \$5.25.
- The regulation of businessmen.* Robert E. Lane. New Haven, Conn.: Yale University Press, 1954. Pp. 144. \$3.75.
- Job evaluation methods.* Second Edition. Charles Walter Lytle. New York: Ronald Press Company, 1954. Pp. 507. \$7.50.
- Teaching tips: A guidebook for the beginning college teacher.* Second Edition. Wilbert McKeachie and Gregory Kimble. Ann Arbor: The George Wahr Publishing Co., 1953. Pp. 108. \$1.50.
- The origins and history of consciousness.* Erich Neumann. New York: Bollingen Series, 140 East 62nd Street, 1954. Pp. 493. \$5.00.
- Psychoanalysis and the education of the child.* Gerald H. J. Pearson. New York: W. W. Norton & Company, Inc., 1954. Pp. 357. \$5.00.
- Developing management ability.* Earl G. Planty and J. Thomas Freeston. New York: Ronald Press Company, 1954. Pp. 447. \$3.75.
- Measurement in today's schools.* Third Edition. C. C. Ross and Julian C. Stanley. New York: Prentice-Hall, Inc., 1954. Pp. 485. \$6.65.
- The clinical interaction.* Seymour B. Sarason. New York: Harper & Brothers, 1954. Pp. 425. \$5.00.
- Problems of infancy and childhood.* Milton J. E. Senn, Editor. New York: The Josiah Macy, Jr., Foundation, 1954. Pp. 196. \$2.75.
- Social science in medicine.* Leo W. Simmons and Harold G. Wolff. New York: Russell Sage Foundation, 1954. Pp. 254. \$3.50.
- Psychology in teaching.* Henry P. Smith. New York: Prentice-Hall, Inc., 1954. Pp. 466. \$4.95.
- Strengthening education at all levels.* Arthur E. Traxler, Editor. Washington, D. C.: American Council on Education, 1954. Pp. 156. \$1.50.
- Techniques of counseling.* Jane Warters. New York: McGraw-Hill Book Company, 1954. Pp. 384. \$4.75.
- Experimental psychology.* Revised Edition. Robert S. Woodworth and Harold Schlosberg. New York: Henry Holt and Company, 1954. Pp. 948. \$8.95.
- Learning theory, personality theory, and clinical research.* The Kentucky Symposium. New York: John Wiley & Sons, Inc., 1954. Pp. 164. \$3.50.



## The Relationship between Mechanical Aptitude and Proficiency Tests for Air Force Mechanics

Major Thomas L. Wood

*Standards Branch, Hq., USAF, Washington, D. C.*

In September, 1952, the United States Air Force installed a world wide proficiency testing program to assist classification officers in selecting airmen qualified for advancement to higher skill levels. The program included use of written tests, custom built by professional test technicians, to cover over 200 specific Air Force jobs.

Development of the tests was carried out at Headquarters Air Materiel Command (Wright-Patterson AFB), Headquarters Air Training Command (Scott AFB), and Headquarters Continental Air Command (Mitchel AFB).<sup>1</sup> Special units, under the direction of officers with psychological training, were organized to write the tests, score the answer sheets, and provide continuous statistical analysis to improve successive forms of each test. Subject matter specialists were selected by the major field commands of the Air Force from master sergeants with wide experience in or supervising the specific jobs for which the tests were built. On the average, five master sergeants worked with a professional test development technician in writing each test. Specialty descriptions of the Airman Career Program were used as guides in building test outlines and in weighting the task elements of each job.

### Problem

Several Air Force studies in the past have shown the relationship between aptitude scores and success in training to be significant (3, 4). Brown and Ghiselli in a sum-

<sup>1</sup> These three units were consolidated into the 2200th Test Squadron, Mitchel AFB, Long Island, N. Y., in April 1953.

mary of the findings of research studies since 1919 concerning the predictive power of tests of intelligence, speed of perception, and spatial and motor aptitudes found aptitude tests to be very useful in predicting training success (1). However, when the aptitude tests are related to job proficiency measures such as speed and amount of production, achievement tests, and supervisor's ratings the predictive power drops considerably.

Since the Air Force had never used job proficiency tests as qualification standards before 1952, it was considered necessary to determine the relationship between the new proficiency tests developed by airman specialists and the Airman Classification Battery (ACB) (2). Aptitude scores from the ACB are used at Air Force Military Training Wings to determine the initial classification and assignment of airmen to technical training courses.

### Rationale

Since the proficiency tests were being used to measure mandatory job knowledge minimums requisite to award of higher skills, they were considered to be a practical criterion of knowledge needed to be successful on the job. An airman who fails to acquire the required minimum knowledge of his job is restricted in skill advancement and, consequently, in promotion to higher rank. Passing the appropriate proficiency test then becomes one objective index of success on the job.

During September, 1952, 9,234 airmen were tested on the senior aircraft mechanic's test. A random sample of 461 cases was selected

to study the relationship between aptitude scores and proficiency scores. The data were divided into four cells on the basis of pass-fail<sup>2</sup> on the proficiency test and qualified-not qualified on the mechanical aptitude test.

It will be noted in Table 1 that 36 (8%) of the airmen failed to attain a qualifying score on the aptitude test, while 115 (25%) of the airmen tested failed the proficiency test. Of those failing the aptitude test, 33 (92%) also failed the mechanic's proficiency test.

Table 1

Relationship between Performance of Aircraft Mechanics on Mechanical Aptitude and Proficiency Tests

	Proficiency Test		Total
	Failed	Passed	
Mechanical Aptitude Test	Passed	82 (18%)	343 (74%)
	Failed	33 (7%)	3 (1%)
	Total	115 (25%)	346 (75%)
			$r_p = .61^*$

Source: Air Force sample, N = 461, tested September 1952.

\* Significant at the 1% level of confidence.

The Pearson  $r$  between the two tests was found to be .61, significant at the 1% level of confidence.

From an Air Force population of 2,426 airmen tested in November, 1952 on the senior vehicle mechanics' proficiency test, a random sample of 303 airmen was selected.

As indicated in Table 2, 23 of 39 (59%) airmen who were below standards in mechanical aptitude passed the proficiency test while only 16 (41%) failed the proficiency exam. In this case it is obvious that me-

<sup>2</sup> Passing on the Proficiency Tests was established at a Standard Score of 80, based on the total Air Force population tested (Standard Score distribution mean 100, std. dev. 20). Aptitude minimum scores are established for each Career Field as a result of research conducted by the Human Resources Research Center to determine aptitude scores predictive of success in technical training courses (5).

Table 2

Relationship between Performance of Vehicle Mechanics on Mechanical Aptitude and Proficiency Tests

	Proficiency Test		Total
	Failed	Passed	
Mechanical Aptitude Test	Passed	26 (9%)	238 (79%)
	Failed	16 (5%)	23 (7%)
	Total	42 (14%)	261 (86%)
			$r_p = .35^*$

Source: Air Force sample, N = 303, tested November 1952.

\* Significant at the 1% level of confidence.

chanical aptitude is not so highly related to the ability to pass a proficiency test custom-built to fit the job. The Pearson  $r$  in this case was .35, significant at the 1% level of confidence.

A further random sample of 189 airmen was drawn from 1,079 senior weapons mechanics tested in November 1952. Table 3 shows that 18 men (10%) were below the minimum in mechanical aptitude. Of these only seven (39%) passed the proficiency test,

Table 3

Relationship between Performance of Weapons Mechanics on Mechanical Aptitude and Proficiency Tests

	Proficiency Test		Total
	Failed	Passed	
Mechanical Aptitude Test	Passed	36 (19%)	135 (71%)
	Failed	11 (6%)	7 (4%)
	Total	47 (25%)	142 (75%)
			$r_p = .35^*$

Source: Air Force sample, N = 189, tested November 1952.

\* Significant at the 1% level of confidence.

while 11 (61%) failed on the proficiency score. The Pearson  $r$  was .35, significant at the 1% level of confidence.

### Summary and Conclusions

Table 4 summarizes the data concerning aircraft mechanics, vehicle mechanics, and weapons mechanics. When each group is divided into a dichotomy of high aptitude and low aptitude, it can be readily seen that the failure rates for the low aptitude men are much higher on the appropriate proficiency test.

Table 4

Failure Rates on Proficiency Tests for High and Low Aptitude Mechanics

Group	Aptitude	Total N	N Failed	Fail Rate
Aircraft Mechanics }	High	425	82	19%
	Low	36	33	92%
Vehicle Mechanics }	High	264	26	10%
	Low	39	16	41%
Weapons Mechanics }	High	171	36	21%
	Low	18	11	61%
Total }	High	860	144	17%
	Low	93	60	65%

Source: Air Force sample, N = 953, tested September and November 1952.

Since not all airmen attend technical schools where they can be evaluated shortly after they receive aptitude tests, it is important for the Air Force to have aptitude scores which have value in predicting not only success in training but also relative performance after experience on the job.

From the data presented it would seem that present mechanical aptitude tests predict future assimilation of job knowledge to a usable degree.

Received November 9, 1953.

### References

1. Brown, C. W. and Ghiselli, E. E. The relationship between the predictive power of aptitude tests for trainability and for job proficiency. *J. appl. Psychol.*, 1952, 36, 370-372.
2. Gordon, Mary Agnes. Validity of the Airman Classification Battery (AC-1A) for Career Guided Samples. *Research Note Pers 51-13*. Human Resources Research Center, Lackland Air Force Base, Texas, July 1951.
3. Gordon, Mary Agnes. A Method of Establishing Minimum Qualifying Scores for Entrance to Air Force Technical Schools. *Technical Report 52-4*. Human Resources Research Center, Lackland Air Force Base, Texas, November 1952.
4. Gragg, D. B. and Gordon, Mary Agnes. Validity of the Airman Classification-Battery AC-1. *Research Bulletin 50-3*, 2nd edition, September 1951. Human Resources Research Center, Lackland Air Force Base, Texas.
5. *Personnel Evaluation Manual*, Air Force Manual 35-8, 1 July 1953, Department of the Air Force, Washington 25, D. C.



## A Comparative Evaluation of Two Approaches to Job-Knowledge Test Construction<sup>1</sup>

Harry M. Mason<sup>2</sup>

*University of Illinois*

Writers concerned with construction of personnel tests (1, 2, 3, 4, 7) advise somewhat different procedures for selecting and editing job-knowledge or trade test contents prior to tryout, but evidence that alternative approaches result in tests having different relationships to criteria is lacking. Experience in editing test content assembled by teams of expert workers under the guidance of test technicians suggested to the writer that two general approaches to the item writing task are used, one of which may be called the job-requirements approach and the other the job-experience approach. It seemed likely that the test items produced through these two approaches would exhibit corresponding differences in their relationships to criteria of job success.

The present study reports an empirical tryout of three newly constructed tests of job knowledge applicable to airplane and engine mechanics maintaining piston-engined aircraft, and three existing tests from the Airman Classification Battery. Two of the new tests were constructed in accordance with the job-experience approach to test construction, and one in accordance with the job-requirements approach. The existing tests were assigned to the two approaches after examination of their contents. After preliminary studies of the new tests with Air Force inductees and airplane and engine mechanic trainees, all six tests were administered to a group of working airman airplane and engine mechanics at an Air Force Base. Results have been evaluated to show the degree to which tests assigned to each approach are

consistent in their relationship to criteria, and differences between the patterns of relationships to criteria characteristic of tests assigned to the two approaches.

### The Two Approaches

The job-requirements approach strives to measure mastery of formally stated job requirements. Tests resulting from this approach are essentially examinations over training courses, job handbooks and similar materials. The job-requirements approach seems likely to dominate production of test items whenever rapid production or revision of tests is required, since it allows the item writer to capitalize upon the organization already existing in published materials.

The job-experience approach to test construction attempts to measure mastery of one or more topics representing distinctive learning opportunities afforded by a job. Test items may relate to what the worker does, to what he may expect to happen on the job, or to knowledge resulting from job aspects having no suspected importance. Since this approach requires the test constructor to select and organize learning opportunities offered by the job, it is difficult to use when tests must be produced quickly.

The two approaches differ in underlying assumptions. The job-requirements approach assumes that workers differ in the degree to which they meet "minimum" job requirements, and that any excess over the minimum level of this type of knowledge results in enhanced job performance. The job-experience approach assumes that all workers retained on the job exhibit above-minimum job knowledge, but that the quality of the worker's adjustment to the job is best reflected in the use he makes of learning opportunities the job affords.

If job requirements were completely realistic, artificial restrictions upon entry to the job and temptations to leave it negligible, and training for job advancement completely relevant, tests resulting from the two approaches might be highly similar. In anything less than the ideal condition, however, differences in tests produced through the two approaches would be expected. In the following paragraphs, tests assigned to the job-requirements approach are called *requirements centered*, and tests assigned to the job-experience approach are called *experience centered*.

<sup>1</sup> This research was supported in part by the United States Air Force under Contract No. AF 33(038)-25726, monitored by the Human Resources Research Center. Permission is granted for reproduction, translation, publication, use and disposal in whole and in part by or for the United States Government.

<sup>2</sup> The writer wishes to acknowledge help and criticism given by Prof. L. H. Lanier.

*Requirements Centered Tests.* Alternate forms of the Electrical Information and Mechanical Principles Tests of the Airman Classification Battery were used.<sup>3</sup> They were assigned to the requirements-centered category because their content and the descriptions given by Gragg and Gordon (5) indicate that they are concerned with principles taught in schools which prepare men to meet requirements for entry into mechanical jobs. The Electrical Information Test contains 30 four-choice items concerned with circuit diagrams and electrical principles. The Mechanical Principles Test presents 15 picture-type items relating to machines encountered in everyday life.

The Training Research Laboratory (TRL) Aviation Mechanics Technical Knowledge Test was constructed for the present study. It consists of 60 four-choice items chosen from data of a long-range study employing 300 job-knowledge test items obtained from Air Force and Navy aviation mechanics schools.<sup>4</sup> Items selected were the best discriminators between airman trainees in early and late phases of technical schools which are prerequisites for entry into the apprentice level of the Air Force job of Airplane and Engine Mechanic. Most of the items relate to the operation or malfunctioning of aircraft components.

*Experience Centered Tests.* An alternate form of the Aviation Information Test of the Airman Classification Battery was assigned to the experience-centered category. It is intended to measure inductees' attempts to gain contact with aviation work. The test contains 30 four-choice items.

The TRL Aviation Information Test was constructed from information contained in Jane (6). It has 30 five-choice items relating to relative cruising speeds and other performance characteristics of well known civilian and military airplanes, names of manufacturers of airplane equipment, and the equipment and components employed in different airplanes. Its content is intended to be continuous in type with that in the Airman Classification Battery Aviation Information Test; it refers to information more readily available to aviation workers than to men outside aviation jobs.

The two aviation information tests were assigned to the experience-centered category because their content may be learned through experience on the job, rather than through attempts to meet formal job requirements through schools.

The TRL Maintenance Techniques Test was made up as a result of interviews with 50 airmen recommended as expert airplane and engine mechanics. Each interview was guided to cover all major airplane systems on the aircraft with which the interviewee was most familiar. Statements of interviewees led directly to test item content, or to intuitive guesses concerning the verbal self-guidance employed by good mechanics in the tasks mentioned. The test contains 76 four-choice items. It emphasizes airframe rather than powerplant systems. Most of the items relate to operations performed by the mechanic, rather than to the mechanical operation of aircraft components.

*Administration of Tests.* Slightly less than four hours were required for subjects to answer the entire battery of six tests and to give personal information and peer ratings. Air Force tests were finished within time limits and were scored according to prescribed formulae; new tests were given without time limits and were scored for number of correct responses.

*Subjects.* Subjects were 204 airplane and engine mechanics chosen as every  $n$ -th name from alphabetical rosters of all airplane and engine mechanics at apprentice to supervisor or technician levels in three airplane maintenance squadrons at Lowry Air Force Base. In sampling,  $n$  was chosen as every third or every fourth man, to give as nearly 70 men per squadron as possible.

*Criteria.* Personal data statements concerning the length of aircraft maintenance experience were the principal criteria employed. In treatment of results, men claiming six years or more of experience are called high-experience men, those with less than six years of experience are called low-experience men. The distribution of experience, with a major mode at less than two years and a minor mode at more than eight years made a breakdown at this point convenient. Division at this point is also in accord with the presumption that the low-experience group is composed primarily of men who have not yet had time to become fully competent, and that the high-experience men have, in general, met whatever effective minimum job requirements exist. High-experience men averaged 10.4 years, low-experience men 1.7 years of experience.

Peer ratings were also employed. Each mechanic ranked for competence the six men whose work habits he knew most thoroughly. Men ranked 3.0 or less, on the average, by four or more peers are called "Good"; those ranked 3.1 or more on the average are called "Poor"; men not ranked by as many as four peers are regarded as "Not Rated." After inspection of test results it was seen that at each experience level, the subgroup rated Poor was different from others, but that subgroups rated Good or Not Rated had essentially the same mean scores on all tests, there being no significant mean differ-

<sup>3</sup> Air Force tests were made available by the Personnel Research Laboratory, Human Resources Research Center, Lackland Air Force Base. Permission to employ these tests is gratefully acknowledged.

<sup>4</sup> The study is being conducted by Dr. E. L. Gaier under the present Air Force contract. Grateful acknowledgment is made for permission to use these items.



Table 1  
Mean Scores of Airplane and Engine Mechanic Criterion Subgroups

Test	Criterion Subgroup*			
	High-Exper. "Poor" (N = 5)	High-Exper. "Other" (N = 50)	Low-Exper. "Poor" (N = 51)	Low-Exper. "Other" (N = 98)
Experience Centered:				
TRL Aviation Information**	14.1	20.7	16.2	15.3
Air Force Aviation Information**	19.3	24.2	21.6	21.1
TRL Maintenance Techniques**	38.9	46.2	41.1	39.4
Requirements Centered:				
TRL Technical Knowledge***	26.1	34.5	30.2	34.6
Air Force Electrical Information	16.1	21.0	19.9	20.7
Air Force Mechanical Principles	10.8	11.0	11.4	11.4

\* Criteria employed were amount of aircraft maintenance experience and peer ratings. See text for method used to establish subgroups.

\*\* Test separates High-experience "other" subgroup from all three remaining groups by a highly significant difference (1 per cent level).

\*\*\* Test separates either "Poor" group from all three remaining subgroups by highly significant differences (1 per cent level).

ences. Consequently, in treatment of results, a subgroup rated Poor is differentiated from one designated as Others at each experience level.

The presumption that the high-experience men are the more competent rests on the assumption that they were at least as able to benefit from experience at the outsets of their careers as were the low-experience men. High-experience men are, as expected, older, 80 per cent being over 30, while less than 7 per cent of the low-experience men are more than 30 years of age. High- and low-experience subgroups alike have 76 per cent who completed high school. Though the older men went to school when educational requirements were lower and thus might have been a more select group, they are also more likely to have achieved their high school graduation through GED examinations. Thus there is no strong evidence of any basic intellectual difference between the two experience levels. Nearly half of the low-experience men had had all their maintenance experience on one type of airplane. None of the high-experience men reported familiarity with less than two types of airplanes. One eleventh of the high-experience men, and one-third of the low-experience men were rated Poor in the peer ratings.

Aptitude Indexes were available for only 117 men, all in the low-experience category. The mean Mechanical Aptitude Index was 6.8, SD 1.33; the mean Technician Specialty Index, comparable to a group intelligence test score, was 6.4, SD 1.88. Neither of these shows a high degree of selection, since Gragg and Gordon (5) found means for both Indexes to be 6.2 in 1,000 presumably unselected inductees in 1949. Air Force Specialty Codes assigned the men ranged

from the "3" to the "7" level, but these were not employed as criteria, since they had not all been assigned through the present uniform procedure.

Among the low-experience men, the peer rating of Poor does not appear to be associated with low Mechanical Aptitude, the proportions of 36 mechanics rated Poor and 81 rated Other having Mechanical Aptitude Indexes of 7 or higher being 53 and 55 per cent, respectively. The proportion of Poor having Technician Specialty Indexes of 7 or higher was 33 per cent, while that for Others was 48 per cent. This difference is nearly significant.

## Results

Preliminary studies had shown that all tests distinguish significantly between groups of inductees and the working mechanic group.<sup>5</sup> In the present study, means of criterion subgroups are compared to determine (a) whether or not a common pattern of subgroup mean differences characterizes the tests assigned to each major approach; and (b)

<sup>5</sup> In preliminary studies, 108 airman inductees and 206 airplane and engine mechanic trainees at the end of training preparatory to the apprentice level of the job were tested. Scores made by these groups indicated that all newly constructed tests differentiate significantly between inductees and trainees, and between trainees and the mechanics tested in the present study. Findings of Gragg and Gordon (5) compared with the mean scores on Air Force tests used in the present study indicate that the Air Force tests likewise distinguish working mechanics from inductees.



whether or not the pattern of criterion subgroup means produced by tests assigned to one approach differs meaningfully from the pattern produced by the tests assigned to the other. Brief consideration is given to correlations between tests, and between tests and Aptitude Indexes.

As shown in Table 1, all of the experience-centered tests show the same rank order of subgroup means. High-experience mechanics rated "Other" have highest scores, followed by low-experience mechanics rated Poor, low-experience Other and high-experience Poor. For each test the difference between the high-experience Other and remaining subgroups is significant beyond the one per cent level, while differences separating the remaining subgroups one from the other are not significant.

Requirements-centered tests do not present a completely consistent pattern of subgroup means, but if the Mechanical Principles Test, which is relatively short and unreliable, is excluded, and fractional mean differences are ignored, a consistent pattern is evident. Both the Technical Knowledge and the Electrical Information tests place the low-experience Other subgroup and the high-experience Other subgroups first, low-experience Poor next, and high-experience Poor last. Both of the subgroups rated Poor are reliably lower in mean score than either of the Other subgroups on the Technical Knowledge Test. No significant differences were produced by the other two requirements-centered tests.

The high-experience Poor subgroup ( $N = 5$ ) occupies bottom position in all tests; it therefore does not assist in differentiating between the two approaches. The most probable explanation for the low test scores of this subgroup is that it represents a recognized minority of poor mechanics who manage to keep their association with the job and the Air Force in spite of limited ability or poor motivation. Three of the five men were ranked last unanimously by all mechanics who listed them among their closest working acquaintances. The characteristics of tests adequate to isolate members of this subgroup do not appear to be critical.

Experience-centered tests discriminate un-

ambiguously in favor of high-experience mechanics generally approved by their peers. Requirements-centered tests, to the extent that they discriminate among the subgroups, tend to isolate the subgroups rated Poor, but do not distinguish knowledge presumably learned primarily through experience. If the experience-centered tests measured nothing except length of tenure on the job, they would have little practical utility. To determine whether or not some mechanics gain this knowledge quickly, a check was made to see how many of the low-experience mechanics scored above the mean of the high-experience Other subgroup, and to examine their personal data and rating characteristics. A total of 28 low-experience mechanics for whom aptitude indexes were available scored above the mean of the high-experience Other subgroup on either the TRL Aviation Information Test or the TRL Maintenance Techniques Test, or both. The breadth of experience represented was substantially the same as that of other low-experience men. Slightly, but not significantly more of the 28 men were rated Poor than among low-experience men generally. The group was outstandingly high on the Technical Knowledge Test, 21 of the 28 having scores of 35 or higher. All but two of these men had both Mechanical and Technician Specialty Aptitude Indexes of 6 or higher, their mean Mechanical Index being 8.0 and their mean Technician Specialty Index being 7.6. Both means are significantly higher than the means of the remainder. For 11 of the 28 men, Mechanical Index was higher, for 15, both were the same, and for two, the Technician Specialty Index was the higher. Thus it appears that experience-centered knowledge may be mastered early in an airman's career, and that high aptitudes are indicative of ability to master it. On the other hand, these men's peers are not inclined to regard them as outstanding mechanics.

Low-experience men rated Poor have a slight tendency to score above other low-experience mechanics on experience-centered tests. As indicated earlier, the Mechanical Aptitudes of the low-experience mechanics rated Poor are substantially equal to those of the Others, but their Technician Specialty

Indexes are nearly significantly lower. The low status of Poor subgroups on the Technical Knowledge Test may be due to this test's higher relationship to general intelligence, indicated by its correlation with aptitudes shown in Table 2. The low-experience Poor subgroup's better status on the experience-centered tests could not be due to better measured mechanical aptitude. It might possibly be due to more realistic attitudes toward the job, resulting in effective but not spectacular job adjustment. Since all three differences are small and non-significant, they could be due to chance.

Table 2 gives means and SDs of tests, split-half reliabilities, average correlation of each test with other tests, and correlation of each test with Mechanical Aptitude Index and Technician Specialty Index, for the 117 low-experience mechanics for whom aptitude indexes were available. Correlations presented indicate the TRL Aviation Information Test to be more independent of other tests and aptitude indexes than any other test used. The Air Force Aviation Information Test has correlations more nearly like those of the requirements-centered tests. The Air Force test is somewhat too easy for working mechanics, having been designed for inductees; this test's ability to distinguish between working mechanics may be based to a considerable

extent upon differences in reading ability, rather than aviation information. On the whole, the experience-centered tests are less highly correlated with aptitude indexes than are the requirements-centered tests. Since these tests are related to criteria among both low-experience and high-experience mechanics, it would appear to be worthwhile to experiment with a mechanical aptitude index depending somewhat more upon tests of knowledge spontaneously acquired before entering the Air Force. It should be borne in mind, however, that the discussion relating to correlations is based only upon differences in their size, without regard for significance of these differences. More studies are needed before a firm interpretation of these relationships is attempted.

A limitation applying to all results of the present study is that the study is cross sectional; differences between mature and less-experienced mechanics could possibly be due to selective attrition. This is not likely, considering the personal data differences between the two experience levels, but only longitudinal studies, employing broad batteries of tests, can establish surely the changes in job-knowledge which differentiate between mature workers who have demonstrated satisfactory adjustment to the job and beginners

Table 2  
Test Score Distribution Characteristics (N = 204)

Test	Mean	SD	Rel.*†	Correlations†		
				Ave., with Other Tests	Mech. Apt.**	Tech. Sp. Aptitude**
Requirements Centered:						
TRL Technical Knowledge	33.2	7.2				
AF Electrical Information	20.5	5.0	81	48	50	42
AF Mechanical Principles	11.3	2.5	75	43	58	42
Experience Centered:						
AF Aviation Information	22.0	4.5	50	29	56	61
TRL Aviation Information	16.7	5.1	71	38	51	48
TRL Maintenance Techniques	41.3	7.4	74	30	23	34
			72	40	58	37

\*Odd-even reliability, corrected by Spearman Brown Formula.

† Decimal points are omitted from correlation coefficients.

\*\* Correlations with Aptitude Indexes are for 117 low-experience mechanics for whom Aptitude Indexes were available.

whose aptitudes are known only in terms of test scores.

The most probable explanation for the different relationship to criteria shown by experience-centered and requirements-centered job-knowledge tests is that requirements-centered tests emphasize formal requirements some of which are not functionally effective, and that for the proven formal requirements measurable with tests, critical levels of mastery have not been incorporated into test materials. Since some discrepancies between stated requirements and learning opportunities accepted by good workers on the job may be expected always to exist, there appears to be a continuing need for empirical studies aimed at improving the coverage of job knowledge tests.

*Received December 15, 1953.*

## References

1. Adkins, Dorothy C. *Construction and analysis of achievement tests*. Washington: U. S. Government Printing Office, 1947.
2. Cronbach, L. J. *Essentials of psychological testing*. New York: Harper Brothers, 1949.
3. Flanagan, J. C. Critical requirements: a new approach to employee evaluation. *Personnel Psychol.*, 1949, 2, 419-425.
4. Flanagan, J. C. The use of comprehensive rationales in test development. *Educ. Psychol. Measmt.*, 1951, 11, 151-155.
5. Gragg, D. B. and Gordon, Mary Agnes. *Validity of the Airman Classification Battery AC-1*. Research Bulletin 50-3, Personnel Research Laboratory, Human Resources Research Center, Lackland Air Force Base, 1951.
6. Jane, F. T., et al. *Jane's all the world's aircraft*. London: Sampson, Low, Marston, 24, 1934 to 52, 1951.
7. Thorndike, R. L. *Personnel selection test and measurement techniques*. New York: John Wiley and Sons, Inc., 1949.



## Retest-Reliability by a Movie Technique of Test Administrators' Judgments of Performance in Process<sup>1</sup>

Arthur I. Siegel

*Institute for Research in Human Relations, Philadelphia*

In work sample performance testing, judgments are often made by the test administrator regarding the manner in which an examinee performs the components of a specific task. For instance, in a Drill Point Grinding Performance Test, the test administrator is asked to make judgments on the manner in which the examinee holds the drill while grinding, on whether the examinee wears loose clothing that could snag in the grinding wheel, on whether the examinee inspects the grinding wheel for cracks prior to grinding, on whether the examinee oscillates the drill while grinding, etc. One of the problems connected with this type of judgment is that the perceptions of the test administrators themselves may vary from time to time and thus may represent an uncontrolled variable in work sample performance testing. Clothing may be perceived on one day as being loose enough to snag in a grinding wheel while two weeks later the same clothing, worn in the same manner, may be perceived as being perfectly safe. Of course, one way in which this type of variation may be partially controlled is to keep the items to be judged gross enough and objective enough so that misperception is minimized. If a definite frame of reference is written into each element scored, and if the observations required are kept gross, then the danger of perceptual variability in examiners may be minimized.

It is still incumbent on the test user to ascertain the reliability of the observations of the people who act as test administrators. The ideal method for determining the consistency of an individual examiner is the situation in which the examinee's performance

is held constant over two separate occasions and the examiners' perceptions allowed to vary. Since the stimulus configuration remains constant, any unreliability shown can then be attributed to variation within the examiner. However, unfortunately, no one can possibly perform the same job in exactly the same manner on two separate occasions. One method by which performance may be held constant is to take a motion picture of the examinee performing the job. The motion picture may then be shown on two separate occasions and the examiner asked to score the motion picture rather than to score the actual live performance. Thus the stimulus situation is held constant over the two time intervals and any variation shown may be attributed to variation within the examiner. Two assumptions of this method are that the movie situation presents the same stimulus configuration to the examiner as does the actual work sample performance test situation and that the examiner scores the movie in the same manner as he would score an actual work sample performance test. Further assumptions are the usual assumptions made in any test-retest reliability check. The principal disadvantage of the motion picture technique is that movies are difficult and expensive to produce. This is especially true for long, involved jobs.

The purpose of this paper is to present our method of using the movie technique for determining intra-examiner reliability and to present the intra-examiner reliabilities obtained from a small group of test administrators.

### Method

A 16 mm. black-white movie was made of a Naval Aviation Structural Mechanic taking a Drill Point Grinding Work Sample Performance Test. The film was unrehearsed and the only instructions given the subject, a randomly selected Aviation Structural Mechanic, were "to grind the drill as he would ordinarily do it." S was told

<sup>1</sup> The data herein reported are a small portion of the data gathered under Contract Nonr-872(00) between the Institute for Research in Human Relations and the Office of Naval Research. The opinions expressed are those of the author and do not necessarily represent the opinions of the Office of Naval Research or of the naval service.

## Care and Use of Tools

1. Did the examinee check the tool rest for proper distance from periphery of the grinding wheel? Yes\_\_\_\_No\_\_\_\_
2. Did the examinee ever adjust the tool rest while the wheel was in motion? Yes\_\_\_\_No\_\_\_\_
3. Did the examinee use a coolant while grinding the drill? Yes\_\_\_\_No\_\_\_\_

## Procedure

4. Did the examinee read the "Examinee Instructions?" Yes\_\_\_\_No\_\_\_\_
5. While grinding, did the examinee oscillate the drill so that heel was moved along the surface of the grinding wheel? Yes\_\_\_\_No\_\_\_\_
6. Did the examinee hold the shank slightly lower than the point while grinding? Yes\_\_\_\_No\_\_\_\_
7. Did the examinee alternate from flute to flute while grinding? Yes\_\_\_\_No\_\_\_\_
8. Did the examinee grind one flute and then the other? Yes\_\_\_\_No\_\_\_\_
9. Did the examinee check the shank of the drill for bends and burns? Yes\_\_\_\_No\_\_\_\_
10. Did the examinee secure the grinding wheel? Yes\_\_\_\_No\_\_\_\_
11. Did the examinee "police up" the work area when securing? Yes\_\_\_\_No\_\_\_\_

## Safety Precautions

12. Did the examinee wear eyeshields or goggles while grinding? Yes\_\_\_\_No\_\_\_\_
13. Did the examinee tap the grinding wheel or check it for cracks prior to its use? Yes\_\_\_\_No\_\_\_\_
14. Did the examinee wear loose clothing or clothing that could snag in the grinding wheel? Yes\_\_\_\_No\_\_\_\_

FIG. 1. Movie evaluation form.

that we were going to take movies of him while he was working. The motion picture cameras and lights were not hidden, but their presence and the knowledge that his behavior was being photographed did not seem to affect the S's behavior.

The motion picture was then first shown in the training room of VC-4, Naval Air Station, Atlantic City, to five Chief Aviation Structural Mechanics. These chiefs had previous experience in work sample performance test administration and were moderately well informed in the general principles of work sample performance test administration. The movie was reshowed to the same chiefs one month after its first administration. One month is usually accepted as a sufficient time interval for forgetting of original responses. Moreover, the chiefs did not know that they would be asked to make exactly the same observations on two separate occasions. Therefore, there was little reason for them to try to remember their original responses.

The chiefs were asked to fill in the Movie Evaluation Form (see Figure 1) during each showing of the motion picture. The Movie Evaluation Form contained fourteen items such as—"Did the examinee check the tool rest for proper distance from the periphery of the grinding wheel?"; "Did the examinee ever adjust the tool rest while the grinding wheel was in motion?"; "Did the examinee wear loose clothing or clothing that could snag in the grinding wheel?"; "Did the examinee check the shank of the drill for bends and burns?"; etc. Sufficient light was allowed in the "theater" so that the chiefs could fill in the forms as the appropriate action was

performed. Thus the motion picture situation was as close as possible to actually scoring a work sample performance test.

## Results

The results in terms of the consistency of the observations of the Chief Structural Mechanics who viewed on two separate occasions the drill point grinding motion picture are presented in Table 1.

Table 1

Intra-Examiner Consistency for Measurements of Performance in Process

Observer	Per Cent Consistency
A	85.6
B	71.4
C	100.0
D	64.3
E	92.8
Mean	82.8

In preparing Table 1, we called S consistent on an item if he answered the item on the second showing of the movie in exactly the same manner that he did on the first showing. Thus:

$$\text{Intra-examiner consistency} = \frac{\text{Number of items answered in exactly the same manner on each showing of motion picture}}{\text{Total number of items on questionnaire}} \times 100$$

The grand mean for intra-examiner agreement was 82.8% with a range from 64.3% to 100%. This mean of 82.8% agreement would usually be considered adequate if converted into a correlation coefficient and interpreted as correlation coefficients are usually interpreted. Of course, these intra-examiner reliability estimates are based on only one motion picture. The danger of generalization from one measure of the reliability of observations of performance in process to all observations of performance in process is self evident.

In view of the range shown, the desirability of determining the reliability of the observations of examiners prior to assigning them to test administrative duties is also indicated. If all examiners show low consistencies, then either the examiner training has been poor or the test itself is inadequate. Naturally, only those examiners with high consistencies are worthy of consideration as test administrators.

The problem of how high examiner consistency must be before it is high enough re-

mains open. A second problem remaining open is that of the effect of increasing the number of judgments to be made on intra-examiner reliability. That is, will the Spearman-Brown prophecy formula hold in this situation?

### Summary

A motion picture technique was described and the results of its use in determining the intra-examiner reliability for performance test administrators' observations of performance in process were indicated. The mean intra-observer consistency for observations of elements of a drill point grinding task on two separate showings of the movie was adequate. However, the range of consistency was great enough to warrant a recommendation in support of careful investigation of the intra-examiner reliability prior to the administration of work sample performance tests.

*Received November 25, 1953.*



## Influences on Merit Ratings

Aaron J. Spector

*Officer Education Research Laboratory, Air Force Personnel and Training  
Research Center, Maxwell Air Force Base \**

Many sources of errors in merit ratings are well known to users of these devices. Laboratory and field investigations have identified errors which may be classified as: (a) characteristic biases of classes of raters, e.g., men, women, peers, etc.; and (b) universal errors, e.g., halo effect, error of central tendency, etc.<sup>1</sup> Somewhat neglected is the fact that the stimulus, the ratee's behavior, may contribute errors which are not ordinarily considered. His total behavior is complex and includes some behaviors which are pertinent and some which aren't pertinent to the factors on which he is being rated. Evaluation of the pertinent behaviors independently of all others may require special training of the raters. This may be especially true when the factors being evaluated are in themselves complex and subjectively loaded, e.g., potentialities of the ratee, cooperativeness, quality of work, etc. Irrelevant characteristics may be so influential as to seriously bias the evaluations on the desired characteristics. The research presented here has been designed to investigate the effects of irrelevant ratee behaviors on ratings assigned to him.

A ratee characteristic, which is irrelevant to the others being evaluated, has been experimentally varied in order to measure its effects on the pertinent characteristics. The variable being manipulated is that of amenability to suggestions. This variable was selected because of the prevalence in industry of situations where suggestions may be accepted or rejected by the ratee and may, therefore, influence the rater's evaluation of

other characteristics. In order to complete the experimental design a second variable, the rater's opportunity to make suggestions to the ratee, was also manipulated.

### Procedures

In five sections of a General Psychology course<sup>2</sup> a guest lecturer was introduced to the class as a student who was interested in becoming a college teacher. The classes, ranging in size from 19 to 30 students, were advised that they would be asked to evaluate his teaching ability after he had lectured. In all classes he delivered the day's lecture in exceedingly poor fashion, making several glaring pedagogical errors, although the material itself was adapted from a well known textbook.<sup>3</sup> After the first 15 minute period, the experimental variable was introduced according to the plan shown in Table 1. Three of the groups (A, B, and C) wrote notes to the lecturer after the first 15 minutes, suggesting improvements to be made in his techniques. A second 15 minute lecture followed, which was as poor as the first. At the conclusion of this lecture the students evaluated the lecturer using a rating scale described below.

After looking over the notes in Group A the lecturer *accepted* the suggestions by thanking the students for them and expressing his intention of modifying his techniques, as per their suggestions. In Group B he *rejected* their suggestions by telling them he had his own ideas on improvement. Although the students in Group C also wrote notes they were *not submitted* until the conclusion of the second 15 minute period of lecture. At this time they made their evaluations and then submitted their suggestions.

\* The author was a member of the faculty at the University of Massachusetts when this study was conducted. He wishes to express his gratitude to his colleagues who contributed their class time to this research, and to Mr. Churchill Morgan for the preliminary analyses of the data.

<sup>1</sup> For a summary of the major studies, see (1). Mahler's (2) review is more comprehensive and recent.

<sup>2</sup> The subjects were sophomore students at the University of Massachusetts. Sections of this course were randomly assigned to the experimental treatments.

<sup>3</sup> The guest lecturer was trained for approximately seven hours in order to insure that his delivery would be comparable in all classes.

Table 1  
Experimental Design

Treat- ment	15 minutes	10 minutes	15 minutes	10 minutes
A	lecture	suggestions written and <i>accepted</i>	lecture	rating
B	lecture	suggestions written and <i>not accepted</i>	lecture	rating
C	lecture	suggestions written but <i>not submitted</i>	lecture	rating
D	lecture	<i>no suggestions;</i> announcement read instead	lecture	rating
E	lecture	<i>no suggestions;</i> ratings made		

Groups D and E were not given the opportunity to suggest any changes to the ratee. Instead of writing suggestions Group D listened to an announcement read by the officially assigned instructor; the amount of time required for the announcement was roughly equivalent to the time other groups used in writing suggestions.

Group E made *no suggestions* and evaluated the lecturer after the first 15 minute period.

The lecturer was evaluated on a rating scale containing five questions measuring: (1) manner; (2) ability; (3) knowledge; (4) potential; and (5) poise. For each question the individual subjects checked one of seven boxes which were ordered on a continuum, as illustrated by question 1, which read, "Compared to others, this lecturer's *manner* while lecturing was: As poor as any I've seen; Considerably worse than most; Not quite as good as most; As good as most; Somewhat better than most; Considerably better than most; As good as any I've seen." The responses on each factor were weighted 0-6, higher scores being assigned to the more favorable responses.

### Results

The most favorable ratings on all five factors were recorded by the *acceptance* group (A), as shown in Table 2. The poorest ratings were given by Group E, which made no suggestions and had only 15 minutes of lecture. The other *no-suggestion* group (D) also rated the lecturer relatively unfavorably. The Mean ratings of B and C groups were equal, but higher than either D or E. It ap-

Table 2  
Means and Standard Deviations of Ratings on Each Characteristic for Each Treatment

Treat- ment	Questions					N	Mean <sub>M</sub>	SD <sub>row</sub>
	1 manner	2 ability	3 knowledge	4 potential	5 poise			
A	M	2.11	2.16	2.95	3.47	19	2.61	.86
	SD	.45	.59	.51	.88			
B	M	1.28	1.52	2.64	2.60	25	1.88	1.19
	SD	.77	.94	.93	1.10			
C	M	1.30	1.64	2.25	2.57	28	1.88	.99
	SD	.70	.98	.95	1.01			
D	M	1.53	1.69	2.29	2.06	35	1.76	.92
	SD	.89	.69	.81	.89			
E	M	1.50	1.13	2.07	1.93	30	1.55	1.34
	SD	1.51	.85	1.18	1.26			
	SD <sub>col</sub>	1.02	.91	.98	1.12			

pears that expression of criticism of the lecturer, via written suggestions, resulted in raters giving higher evaluations than when the raters had no opportunity for this expression. These results obtained when the rater's suggestions were not submitted to the ratee, as well as when they were submitted and accepted or rejected.

The most favorable ratings, however, were consistently made by the group whose suggestions were accepted by the ratee. Apparently, amenability to suggestions or expressed intention of compliance with the suggestions, operated to bias the raters' evaluations of the lecturer.

The data were analyzed further by analysis of variance.<sup>4</sup> An F ratio, obtained with *total scores*<sup>5</sup> of all subjects in each treatment, indicated that the mean total scores were significantly influenced by the treatments accorded the groups (Table 3).

Table 3

Analysis of Variance of Total Merit Rating Scores of all Subjects in Five Experimental Treatments

Source of Variation	df	M.S.	F	P
Between treatments	4	2123.66	5.49	.01
Within treatments	132	386.51		

The ratings on each characteristic were then examined. F ratios indicated that the responses on four of the questions varied significantly between groups (Table 4). That is, the experimental treatments accorded to the groups differentially affected their ratings on four out of the five characteristics.

The only Between Groups variance which was not significantly different from chance was on evaluation of the lecturer's ability. If the students measured the lecturer's ability by the amount they had learned or by the quantity of notes they could take, it is understandable that their evaluations would agree since neither learning nor note taking came easily from his lecture.

<sup>4</sup> The variances were found to be homogeneous by Bartlett's test.

<sup>5</sup> An average intercorrelation of .18 was obtained between items on the rating sheet, using Peters and Van Voorhis' formula (4, pp. 196-200).

Table 4

Analysis of Variance of Ratings on Each Question for All Treatments Simultaneously

Question	Between Treatments		F	p
	Within Treatments	df		
1	2.902	4	2.76	.05
	1.052	132		
2	.025	4	.03	
	.801	132		
3	2.882	4	2.60	.05
	1.109	132		
4	8.516	4	6.50	.01
	1.310	132		
5	3.534	4	2.61	.05
	1.356	132		

However, no such simple criteria existed for rating his manner, knowledge, poise, or particularly his potential. These ratings may reflect personal frames of reference and hence are more readily influenced by extraneous factors such as acceptance or rejection of suggestions. Similarly, the factors of promotability and quality of work, which are frequently found on industrial merit rating scales, may be especially prone to the influence of irrelevant behaviors of the ratee.

### Discussion

The cathartic effects of expression of criticism via written messages, noted above, are consistent with the findings of Thibaut and Coules (4). Their data indicated that persons who were insulted, and then allowed to express their hostility toward the instigator, via written notes, later made a greater number of friendly remarks about the instigator than did other insulted persons who had no opportunity to express their hostility. The present data suggest that poor impressions, like ill feelings, may be altered or reduced, by their expression.<sup>6</sup> Low ratings may re-

<sup>6</sup> The dynamics of this phenomenon are described by Newcomb (3) in his discussion of "autistic hostility."



flect a barrier in communications between the supervisor and his subordinates, rather than true deficiencies of the ratees. Therefore, a likely hypothesis is that merit ratings in industry may be influenced by the degree to which the rater feels free to criticize or make suggestions to ratees.

The practical importance of the finding that irrelevant characteristics of ratees may bias raters' judgments is difficult to evaluate without more knowledge of: (a) the kinds of ratee behaviors which act in this way; and (b) the amount of bias these behaviors induce. At any rate, it is clear that amenability to suggestions induces sufficient bias to significantly affect ratings on several factors.

### Summary

Students in five sections of a general psychology course listened to a lecture which was intentionally delivered in poor fashion. They were then asked to rate the lecturer on five characteristics, using a seven point scale. Before they rated him three of the groups suggested methods by which the lecturer might improve his techniques. One of these groups *did not submit* their suggestions to the lecturer; in another group the lecturer *rejected* the suggestions, while in the third he *ac-*

*cepted* them. In two other groups the subjects *did not write* suggestions. In no case did the lecturer actually implement the suggestions, or improve his delivery.

The ratings were: (a) consistently most favorable in the *acceptance* group; (b) more favorable in the *suggestion* than the *no-suggestion* group; (c) significantly different on the characteristics of manner, poise, potential and knowledge.

It has been suggested that poor ratings may reflect barriers in communications between the rater and the ratee, rather than true deficiencies in the ratees.

*Received November 20, 1953.*

### References

1. Guilford, J. P. *Psychometric methods*. N. Y.: McGraw-Hill, 1936.
2. Mahler, W. R. *Twenty years of merit rating, 1926-1946*. N. Y.: The Psychological Corp., 1947.
3. Newcomb, T. Autistic hostility and social reality. *Hum. Rel.*, 1947, 1, 69-86.
4. Peters, C. C. and Van Voorhis, W. R. *Statistical procedures and their mathematical bases*. N. Y.: McGraw-Hill, 1940.
5. Thibaut, J. W. and Coules, J. The role of communications in the reduction of inter-personal hostility. *J. abnorm. soc. Psychol.*, 1952, 47, 770-778.

## Psychological Research on Accidents: Some Methodological Considerations<sup>1</sup>

Kenneth S. Teel

*Human Factors Operations Research Laboratories*

and

Philip H. Du Bois

*Washington University*

Recently controversy has arisen over the methods that can appropriately and meaningfully be used in psychological research on accidents. An article by Mintz and Blum (8) advocating use of the Poisson distribution and analysis of variance for estimating the extent of personnel-centered accident liability precipitated the discussion. In opposition, Maritz (7) argued that "... the direct technique of 'correlating consecutive periods' is indispensable." More recently, however, Blum and Mintz (1) and particularly Webb and Jones (13) have pointed out that the different techniques are basically the same and should for all practical purposes yield similar results.

This paper aims to point out certain shortcomings of these methods and to propose more refined solutions.

### The Poisson Distribution

A frequency distribution of numbers of accidents by individuals can be symmetrical only when the mean number of accidents is appreciably greater than unity. When there are fewer accidents than individuals, the zero category must have the greatest frequency, and superficially at least the obtained distribution must resemble the Poisson distribution—often used to estimate the extent to which variations in accident histories may be attributable to chance factors. Methods of computing the theoretical distribution and of testing the difference between it and the one

actually obtained are explained elsewhere (2, 8, 9).

Interpretation of the obtained results is, however, fairly difficult. First, let us consider the simpler of the two possibilities—that in which the obtained distribution deviates significantly from a Poisson. Ordinarily this result is interpreted as indicating the presence in the population of varying degrees of accident proneness. Such an interpretation is justified only if all persons were exposed to the same hazards; if this were not the case, the significant deviation from the Poisson might be reflecting little more than differences in exposure.

Second, let us consider the opposite result—that in which the obtained distribution does not deviate significantly from a Poisson. This result is usually interpreted as indicating that chance factors may account for the obtained variations in accident records and that the null hypothesis cannot therefore be rejected. Here again, however, the interpretation is open to question, for representation of the data by a Poisson does not eliminate the possibility of significant correlation, either between accident records in successive periods or between accident records and logically related predictors. Maritz (7) has already demonstrated the possibility of obtaining a correlation of .80 between two Poisson distributions of accidents in separate time periods.

The coarseness of the measuring unit—the fact of an accident—presents further logical problems in interpreting the results of a Poisson fit. For administrative purposes, some arbitrary definition of an accident is necessary; however, rigid adherence to the

<sup>1</sup>This paper was prepared at MacDill Air Force Base while PHDB was acting as consultant to the Human Factors Operations Research Laboratories. The opinions expressed in this paper are those of the authors and do not necessarily reflect the views of the Air Force.

definition forces into a discrete series behaviors which really are on a continuum. It seems perfectly reasonable to assume that the same behaviors which might in one instance result in extensive materiel damage or personal injury might in another result in merely a "close call." It therefore seems logical to assume also that the persons comprising each of the accident frequency groups (zero, single, or multiple) exhibit these behaviors in varying degrees and could, if more complete information were available, be further differentiated.

In short, use of the Poisson distribution is at best a preliminary step in the study of accident proneness. It serves only to provide a quantitative estimate of the stability of accident rates. It furnishes no information whatsoever on the relationships between specific personnel and situational factors on the one hand and accidents on the other.

### Correlational Techniques

Correlational techniques have most frequently been used in accident research to provide a quantitative estimate of the consistency of the accident behavior of individuals from one period of time to another. In attempts to equate exposure in the two periods, most investigators have chosen periods of odd and even months (4) or odd and even days (12). Obtained correlations are typically low, indicating little reliability of the accident criterion.

The magnitude of the correlation is a function of both the accident rate and the length of the exposure period. Because of the coarseness of the accident criterion used in most studies, accident rates are low and, hence, the exposure periods required for obtaining reliable measures are almost prohibitive in length for most practical purposes.

Interpretation of the results of studies of this sort is difficult because of the lack of adequate information on exposure. Even if the assumption of equal exposure among individuals be granted, the correlation of accidents in separate time periods provides little more than a preliminary estimate of the stability of the accident criterion. Such an approach affords no means of identifying per-

sons most liable to accidents except from their accident histories. It does not, therefore, provide any basis for predicting which members of an independent sample will have accidents until after some have experienced them. Since our goal is not only to reduce but also to prevent accidents, this approach obviously is inadequate.

A more adequate but less widely-used application of correlational techniques in accident research is that of correlating individuals' scores on theoretically-related predictor variables (such as intelligence, psycho-motor ability, physical condition, etc.) with their accident records. In actual practice, however, this approach (3) too has typically yielded low correlations—again probably because of the grossness of the accident criterion.

### Suggested Refinements

The design of research on personnel factors in accidents is admittedly difficult. However, we would like to suggest several refinements which would make such research more useful.

The first and most needed refinement is a more sensitive criterion measure. The systematic collection of information on "near-accidents" and the critical behaviors involved therein would provide such a criterion (10). The Strategic Air Command, USAF, has already adopted a policy of collecting these data and using them as a basis for both remedial and preventive training in how to react in emergency situations. Thus, "near-accident" data may have practical value, even before they accumulate in sufficient quantity to provide a reliable criterion against which personnel factor variables can be validated. The high ratio of near-accidents to accidents indicates that the length of the exposure period required for reliable measurement would be considerably shorter.

A second much-needed refinement is better differentiation between "personal" and "situational" accidents.<sup>2</sup> We must discriminate between those accidents caused largely by situa-

<sup>2</sup> The comments contained in this and the following paragraph apply to whatever criterion is used—be it accidents, near-accidents, or a combination of both. The term "accidents" alone is used in the subsequent discussion solely for simplicity of presentation.



tional factors and those caused primarily by personnel factors. Theoretically, situational accidents, such as those caused solely by materiel failure, occur completely independently of the personnel involved. If this be the case, we should not expect to be able to predict them from a knowledge of personnel factor variables only. The inclusion of situational accidents in a study of personnel factors will lower the obtained correlations and make them extremely difficult to interpret.

One might legitimately ask if the proposed differentiation can successfully be made. The recent study of Kubis, Buckley, and Sackman (5) indicates that it can.

A third requisite for more meaningful studies of personnel factors in accidents is the systematic collection of more complete information on exposure to hazard. Whenever possible, records of the time spent in performance of the various aspects of a job should be maintained, for they provide the basis for determining relative hazards. Once we have adequate data on the time spent and accidents incurred on the several parts of a job, we can determine the risk per unit of time for each. We can then compute an index of exposure for each person which weights his experience in the various phases of the job by the risk associated with each. An index of this sort has recently been developed and successfully applied to an Air Force population by Warren et al. (11).

If such an index be available for all members of our sample, we need not make the questionable assumption of equality of exposure. Instead, we can, to some degree at least, remove the effect of differential exposure from our computations by partial correlation or other appropriate technique.

A fourth proposed refinement is the computation of correlations between individuals' scores on certain theoretically-related predictor variables and their records of accidents in which the measured traits are thought to be important. For example, we would hypothesize that psycho-motor test scores should predict only those accidents in which psycho-motor deficiencies are a primary contributing factor. Testing the significance of the obtained correlation coefficient would then en-

able us to confirm or refute our hypothesis. On the other hand, correlation of scores on such tests with over-all accident records would be likely to obscure any relationships which might exist.

In groups where the number of accidents is small and the members are engaged in diverse tasks for which the degree of hazard is difficult to estimate, correlational studies of the type just outlined are not likely to prove worthwhile. Consequently, a different attack must be made on the accident problem. The one recommended is detailed situational analysis, accomplished by experienced job analysts or safety engineers or both and followed by administrative actions designed to overcome identified hazards. As a matter of fact, a thorough situational analysis (6) is indispensable even in those instances (such as in the military and in large industrial organizations) in which the correlational approach is applicable, for the contribution of personnel factors to existing accident rates is usually considerably less than that of situational variables.

Until those factors, either within individuals or within situations, related to accidents are clearly identified and remedial actions instituted, much research remains to be done. As yet, our research efforts have not approached this point.

Received December 21, 1953.

#### References

1. Blum, M. L. and Mintz, A. Correlation versus curve fitting in research on accident proneness: Reply to Maritz. *Psychol. Bull.*, 1951, 48, 413-418.
2. Burke, C. J. A chi-square test for "proneness" in accident data. *Psychol. Bull.*, 1951, 48, 496-504.
3. Fitzpatrick, R., Vasilas, J., and Peterson, R. *Personnel and training factors in fighter aircraft accidents*. Human Factors Operations Research Laboratories, Bolling AFB, Washington 25, D. C., 1953.
4. Ghiselli, E. E. and Brown, C. W. Accident proneness among streetcar motormen and motor coach operators. *J. appl. Psychol.*, 1948, 32, 20-23.
5. Kubis, J. F., Buckley, E. P., and Sackman, H. *The fatal ground accident in the United States Air Force*. Human Resources Research Laboratories, Bolling AFB, Washington 25, D. C., 1952.

6. LeShan, L. L. and Brame, J. B. A note on techniques in the investigation of accident prone behavior. *J. appl. Psychol.*, 1953, 37, 79-81.
7. Maritz, J. S. On the validity of inferences drawn from the fitting of Poisson and negative binomial distributions to observed accident data. *Psychol. Bull.*, 1950, 47, 434-443.
8. Mintz, A. and Blum, M. L. A re-examination of the accident proneness concept. *J. appl. Psychol.*, 1949, 33, 195-211.
9. Thorndike, R. L. *The human factor in accidents with special reference to aircraft accidents*. USAF School of Aviation Medicine, Randolph AFB, Texas, 1953.
10. Vasilas, J. N., Fitzpatrick, R., DuBois, P. H., and Youtz, R. P. *Human factors in near accidents*. USAF School of Aviation Medicine, Randolph AFB, Texas, 1953.
11. Warren, N. D., Mackie, R. R., Simmons, R. F., and Rodman, I. L. *An index of accident exposure for flying in the USAF*. Human Factors Operations Research Laboratories, Bolling AFB, Washington 25, D. C., 1953.
12. Webb, W. B. and Jones, E. R. *Repeater pilot accidents in the United States Air Force*. Human Resources Research Laboratories, Bolling AFB, Washington 25, D. C., 1952.
13. Webb, W. B. and Jones, E. R. Some relations between two statistical approaches to accident proneness. *Psychol. Bull.*, 1953, 50, 133-136.

## Time Intervals between Accidents

Alexander Mintz

*City College of New York*

This study was undertaken in the hope of contributing to a clarification of the evidence on which the widely accepted theory of individual differences in accident liability is based. This evidence is incomplete. One of the principal facts included in it is the frequently-good fit of obtained accident distributions to the so-called unequal liabilities distribution<sup>1</sup> derived by Greenwood and Yule (4). This derivation includes the assumption of constancy of individual liability to accidents, i.e., the notion that accident liability of particular persons does not change with the passage of time or with the occurrence of accidents.

However, there are theoretical considerations which suggest that such an assumption is not likely to be exact. An accident can be expected to function at times as a traumatic experience and to disrupt subsequent behavior. It can also be expected to function as a punishment and, as such, to have one or another effect on the learning of the individual. There are some accident distributions which do not fit theoretical distributions based on assumptions of the constancy of accident liability (3). Very likely, in such cases accident proneness is affected by the accidents.

Even in the cases in which theoretical distributions based on the assumption of constant accident proneness do fit the data, the possibility of inconstant accident proneness cannot be excluded. It has been shown, e.g., by Irwin (6), that the same distribution which Greenwood and Yule derived in part from the assumption that accident liability varies from one individual to another but is constant for each individual, can also be derived from the assumption that there are no initial variations in accident liability, but that instead each accident increases the proneness of the individual by a constant amount. It can undoubtedly also be derived from an assumption of large initial differences in accident liability and decrease of accident liability with the occurrence of accidents.

<sup>1</sup> Or negative binomial distribution.

Thus only tentative inferences may be drawn about the probable underlying distribution of accident liability from an obtained set of accident records, unless something is known about whether and how accidents occurring to people affect their accident liability. Not even the existence of initial differences in accident liability in the group may be inferred with certainty without such knowledge. In the absence of such knowledge, the assumption of unchanged liability after accidents was generally either implied (1) or explicitly made (10, 11) by workers in this field.

The factual evidence on the validity of the assumption of accident liability as unchanged by accidents is very scanty. Irwin has commented upon a few results on accident rates of groups of people in consecutive periods which were opposed to his hypothesis of accident proneness increasing with accidents. The accident rates did not tend to increase; the changes were slight, and, if anything, the rates tended to drop. A rather similar finding has been discussed by Kerrich (8). On the other hand, Horn has presented material on time-intervals between airplane accidents which suggested to him that accident susceptibility is temporarily increased by accidents. He recommended adjustment techniques for pilots following accidents. Thus the question may be of great practical importance to those interested in preventing accidents.

### The Problem

It was thought that further research on time-intervals between accidents was desirable. Increasing accident proneness should show itself as a trend toward decreased time-intervals between accidents, while decreasing accident proneness should show the opposite trend. The problem of this paper was to discover whether trends towards such changes of time intervals do or do not occur.

However, there are methodological difficulties in such research. Thus it is not immediately obvious, how the interval before the first accident and the interval after the



last accident during the arbitrary observation period should be treated, compared to the time intervals between accidents. This paper reports an attempt to deal with one set of data on time-intervals between accidents. It is hoped to provide examples of the type of information which can be obtained from the study of time-intervals, and to present some material relevant to the methodology in this field. The material is examined chiefly in relation to two possible theories: first, that accident proneness is constant for each individual; and, conversely, that proneness is increased with accidents.

### The Data

The data examined were accident records of 178 taxi drivers, made available by Dr. E. Ghiselli, whose cooperation is appreciated. The period covered was one year. For each driver, the weeks in which accidents occurred were indicated. All drivers had worked for the company at least a year prior to the beginning of the observation period. Six drivers who resigned from their jobs during the observation period, or who were absent from work for eight or more weeks, were eliminated. Thus records of 172 drivers were included in this study.

### The Mathematical Background

In order to discover what the time intervals before, between, and after the accidents indicate about the possible effects of accidents on accident liability, it is essential to compare them to the statistical expectancies based on the assumption that accidents are distributed over a time period completely at random. The hypothesis of random distribution of points within an interval has been previously studied by Whitworth (13), Greenwood (2), Moran (12), and Maguire, Pearson and Wynn (9). It assumes that each accident is independent of all other accidents and that its occurrence is equally probable at all times during the period. However, this is assumed only if each accident is viewed as a separate entity, defined in terms of what happens (e.g., sideswiping a particular telephone pole) rather than in terms of the position of the accidents relative to each other. Sideswiping the telephone pole is more likely to be the

first accident if it happens early in the observation period than if it happens late; and so with other types of accidents. The probability that a particular accident is the first one during the observation period is proportionate to the probability that no other accident has yet taken place. This probability decreases with the passage of time.

If  $n$  accidents have happened to an individual during a time interval of unit duration, the probability of the first accident happening at time  $x$  decreases in proportion with  $(1-x)^{n-1}$  as  $x$  increases. The probability function of the first accident within a total time interval of unit length is given by the expression  $n(1-x)^{n-1}$ . The probability of the second accident at time  $x$  involves first, that one accident must have taken place already, and second, that no other accident shall have happened yet. Its formula is  $n(n-1)x(1-x)^{n-2}$ , and similar expressions may be derived for the probabilities of the times of the other accidents.

Probability functions can generally be used for the computation of theoretical means and standard deviations. Such computations indicate that, in terms of the null hypothesis, statistical expectancies for the mean time-intervals from the beginning of the observation period to the first accident; from the first to the second; and so on, including the time-interval between the last accident and the end of the observation period are the same. Similarly, the expectancies of the variances of the time-intervals are also identical. In studying the possible effects of accidents on accident liability, the periods before the first accident and after the last accident may be treated in the same manner as the intervals between accidents.

### Results

Of the 172 drivers included in the computation, 60 had no recorded accidents and 112 drivers had one or more accidents, ranging up to 25. The accident distribution was very different from the theoretical distribution which results from the assumptions of equal and constant accident liability. In an equal liability or so-called Poissonian distribution, the variance of accidents is equal to the mean. In the Ghiselli data, it is about

six times as large as the mean. The distribution seems to be capable of being explained in terms of the hypothesis of large stable differences in accident liability. On the other hand, in accordance with the considerations mentioned earlier, it also can be explained in terms of other assumptions, e.g., that of linear increase of accident liability with accidents.

What do the time-intervals indicate? They were first examined separately for groups of drivers with different accident records.

A total of 45 drivers had one accident each. The theoretical expectancy for the position of the mean time of a single accident is 26 weeks. The obtained mean was 21.9 weeks. The critical ratio was 1.99,<sup>2</sup> which is significant at the .05 level. It should be noted that this suggestive difference is in the opposite direction from the one which would be expected if accident liability increased with accidents. The question was not investigated whether this result was due to the fact that accident liability decreased with the first accident in this group, or whether it was produced by seasonal fluctuations.

In the two-accident case, the situation was somewhat similar. The mean durations of the three time-intervals (up to the first accident, between the first and second accidents, and from the second accident until the end of the observation period) were 13.1; 15.6; and 23.2, respectively. The theoretical expectancy is 17.33, with a standard error of 3.06 weeks. The differences between the time-intervals are again suggestive of a decrease in accident liability after accidents, but the result does not seem to be statistically significant.

The situation was somewhat different in the cases which had three, four, and five accidents. Here the first and last time-intervals were longer than the time-intervals between accidents, but this finding was again of doubtful statistical significance. Groups with more than five accidents were too small for detailed presentation.

The significant fact which emerges from the examination of the mean time-intervals of groups of drivers, classified on the basis of

number of accidents, is that there was no consistent trend toward a decrease of time-intervals with repeated accidents. Table 1 presents the data.

In the preceding discussion, separate comparisons of time intervals were made within each group of drivers with a particular number of accidents. These groups were for the most part very small. Therefore the data were also treated in another way, in terms of cumulative groups. For all drivers who had one or more accidents, the mean time interval before their first (or only) accident was ascertained. The mean time interval before the second accident was ascertained for all drivers with two or more accidents, and for the same group the mean time interval before the first time accident was also computed. Similarly, the mean time intervals before the third accident and before the first accident were determined for the group with three or more accidents, and so on.

The results of these computations are presented in Table 2, the first column of figures giving the mean times between the consecutive accidents and the second column giving the mean times before the first accidents of the same people, and the third column presenting the differences. It should be noted that according to both hypotheses considered in this paper, the figures in the first column should tend to decrease as one proceeds down the table.

According to the theory of individual differences in accident proneness, the same decrease is expected in the second column; this is to be expected because the bottom of the table deals with drivers who had repeated accidents, because repeated accidents are apt to be indicative of high accident proneness, and because high accident proneness is apt to result in short time intervals both before the first accident and between the later accidents. According to the theory of increased proneness following accidents the decrease should be much more pronounced in the first column than in the second one,<sup>3</sup> and the differences in the third column should tend to be negative and to increase in absolute amount.

<sup>2</sup> The critical ratio rather than the t-ratio was used because a theoretical standard deviation could be and was utilized.

<sup>3</sup> There are two reasons for expecting some downward trend in these figures according to the behavior disruption theory. First, there are selective factors: people who had the first accident early "by chance"



Table 1

Mean Times \* Before the First Accident, Between Accidents, and After the Last Accident

One-accident group	(n = 45): 21.9; 30.1
Two-accident group	(n = 16): 13.1; 15.7; 23.2
Three-accident group	(n = 13): 16.1; 9.6; 12.9; 13.4
Four-accident group	(n = 13): 14.0; 8.6; 5.5; 5.6; 18.3
Five-accident group	(n = 4): 9.7; 6.0; 5.5; 7.8; 8.2; 14.8
Six-accident group	(n = 5): 5.9; 4.8; 6.8; 9.2; 10.6; 10.0; 4.7
Seven-accident group	(n = 3): 2.5; 10.0; 5.3; 5.7; 9.0; 7.0; 6.0; 6.5
Eight-accident group	(n = 2): 1.0; 2.0; 6.5; 7.5; 5.5; 2.0; 2.5; 11.5; 13.5
Nine-accident group	(n = 3): 2.2; 2.7; 9.7; 6.7; 3.0; 3.3; 1.3; 8.0; 5.0; 10.2
Eleven-accident group	(n = 2): 8.5; 2.5; 1.5; 2.5; 1.5; 3.5; 6.0; 4.5; 8.5; 2.5; 6.0; 5.5
Twelve-accident group	(n = 1): 2.5; 1; 6; 4; 6; 2; 1; 2; 6; 1; 2; 4; 14.5
Thirteen-accident group	(n = 1): 1.5; 1; 5; 1; 2; 1; 1; 8; 2; 13; 1; 11; 4; 0.5
Fifteen-accident group	(n = 1): 1.5; 1; 1; 3; 3; 2; 3; 6; 5; 1; 4; 5; 3; 10; 2; 1.5
Sixteen-accident group	(n = 1): 2.5; 7; 2; 1; 1; 1; 5; 6; 3; 1; 5; 2; 6; 1; 2; 3; 3.5
Eighteen-accident group	(n = 1): 0.5; 2; 1; 12; 1; 1; 3; 1; 1; 1; 1; 5; 1; 2; 1; 3; 10; 4.5
Twenty-five-accident group	(n = 1): 3.5; 1; 1; 1; 2; 1; 1; 5; 2; 1; 1; 1; 6; 1; 6; 1; 1; 1; 7; 1; 1; 3; 1; 1; 1; 0.5

\* In computing the mean times it was assumed that the accidents occurred in the middle of the week.

The increase of the magnitude of the differences should occur because the accidents intervening between the first accident and the later ones are assumed to increase their accident proneness. This factor would be assumed to produce a marked decrease in the figures of the first column; it would be assumed to be lacking in the case of the figures in the second column, in which only a weaker downward trend would be expected.

There are a number of statistical procedures by means of which the agreement of the two hypotheses with the data could be tested. However, their presentation would have required much space, mainly because of two difficulties: the groups of drivers overlap, so that the figures in the second column are not independent, and the theoretical distributions of the time intervals are not normal. It was not thought that the expected gain from the treatment of the data embodying these considerations was likely to justify the added space. Therefore the material is treated in terms of a simple inspection of the table.

The expected tendency towards decreasing time intervals between the higher numbered accidents is present. As the incidence of ac-

have more time left in which they may have additional accidents; second: the drivers in this study were not new and may have developed differences in accident proneness as a result of accidents occurring before the observation period.

cidents rises, the time interval before the first accident also tends to grow shorter, to about the same extent. As one reads down the table, there is no tendency toward larger negative values of differences. There are some fluctuations in the values of these differences, but these fluctuations are not large, do not suggest an intelligible pattern, and according to tentative computations do not seem to be statistically significant.

### Discussion

These results are clearly not in favor of the hypothesis of increased accident susceptibility with accidents. For this set of data the theory of proneness, varying from person and reasonably-constant for each person appears to be more appropriate.

This conclusion requires qualifications. It should be noted that certain factors were not taken into consideration in this study. The possibility of seasonal fluctuations in accident rates was one such factor. Another factor not considered had to do with the different distances driven by the different drivers. Both of these factors are likely to have functioned as sources of variation in the accident rates, and taking them into account should have given a somewhat better test of the hypothesis of constant accident proneness of individuals.



Table 2

Comparison of Mean Time Intervals in Weeks Before First Accident and Before Later Accidents

	Mean Time Interval	Mean Time Interval of the Same Drivers Before First Accident	Difference
Before 1st accident (112 drivers)	15.1		
Between 1st & 2nd (67 drivers)	8.9	10.7	-1.8
Between 2nd & 3rd (51 drivers)	7.3	9.9	-2.6
Between 3rd & 4th (38 drivers)	6.0	7.8	-1.8
Between 4th & 5th (25 drivers)	6.0	4.5	1.5
Between 5th & 6th (21 drivers)	4.8	3.5	1.3
Between 6th & 7th (16 drivers)	3.3	2.8	0.5
Between 7th & 8th (13 drivers)	6.5	2.9	3.6
Between 8th & 9th (11 drivers)	4.6	3.2	1.4
Between 9th & 10th (8 drivers)	3.2	3.6	-0.4
Between 10th & 11th (8 drivers)	3.2	3.6	-0.4
Between 11th & 12th (6 drivers)	4.0	2.0	2.0
Between 12th & 13th (5 drivers)	4.8	1.9	2.9
Between 13th & 14th (4 drivers)	3.2	2.0	1.2
Between 14th & 15th (4 drivers)	3.0	2.0	1.0
Between 15th & 16th (3 drivers)	1.7	2.2	-0.5
Between 16th & 17th (2 drivers)	2.0	2.0	0.0
Between 17th & 18th (2 drivers)	5.5	2.0	3.5
Between 18th & 19th, 19th & 20th, etc. (1 driver)	7, 1, 1, 3, 1, 1, 1	3.5	3.5, -2.5, -2.5, -0.5, -2.5, -2.5, -2.5

However, this hypothesis is probably only an approximation which is not applicable to all individuals and groups. The records of a

few drivers suggest temporary fluctuations of accident proneness with some individuals. Temporary increases in accident proneness

may well be due to periods of emotional stress.

However, there is need to investigate the statistical significance of such apparent fluctuations in accident proneness or liability. Maguire, Pearson and Wynn (9), Greenwood (2), and Irwin (7) have pointed out certain difficulties in determining the statistical significance of departures of sequences of time intervals from randomness, and it is not entirely clear to this writer whether the problem has been solved.

The apparent lack of systematic effects of accidents on accident rates found in this study need not hold for all groups. It may have partly resulted from the fact that all drivers had worked for the company at least a year before the observation period. Considerations based on the psychology of learning suggest that accident proneness might be less constant with inexperienced workers. Research on time-intervals between accidents for inexperienced workers is worth attempting.

Our finding is not in agreement with Horn's conclusion that accident susceptibility is temporarily increased by accidents. This disagreement with Horn's conclusion may represent a difference between different kinds of accidents, since his data dealt with airplane accidents and ours with accidents to taxi-drivers. On the other hand, the discrepancy may be due to different statistical treatments. Possibly a statistical artifact was involved in his conclusions. Horn's tables showed a relative preponderance of short time intervals over longer ones between consecutive accidents. However, he was apparently not aware of the nature of the distribution of the time intervals between events distributed at random within a period of time, and his tables do not indicate whether or not his results differed from chance expectancy. The matter calls for further investigation.

### Summary

Much of the evidence in favor of the commonly accepted hypothesis of individual differences in accident proneness is only valid if one assumes that accident proneness of individuals is not affected by accidents in which

they are involved. The validity of this assumption is investigated in terms of a study of time intervals between consecutive accidents of a number of taxi-drivers. Some features of the relevant mathematical theory of the random distribution of events in time are reviewed. The findings pertaining to the time intervals between accidents suggest, that, for the group studied, the customary assumption of unchanged accident proneness following accidents is approximately true.

Received December 28, 1953.

### References

1. Cobb, P. W. The limit of usefulness of accident rate as a measure of accident proneness. *J. appl. Psychol.*, 1940, 24, 154-159.
2. Greenwood, M. The statistical study of infectious diseases. *J. Roy. Stat. Soc.*, 1946, 109, 85-109.
3. Greenwood, M. and Woods, H. M. The incidence of industrial accidents upon individuals with specific reference to multiple accidents. *Industr. Fatigue Res. Bd. Rep't* 4, 1919.
4. Greenwood, M. and Yule, G. U. An enquiry into the nature of frequency distributions representing multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents. *J. Roy. Stat. Soc.*, 1920, 83, 255-279.
5. Horn, D. A study of pilots with repeated accidents. *J. Aviat. Medic.*, 1947, 18, 440-449.
6. Irwin, J. O. Discussion on Chambers and Yule's paper. *J. Roy. Stat. Soc.*, 1941, 7 (suppl.), 101-107.
7. Irwin, J. O. Discussion on Professor Greenwood's paper. *J. Roy. Stat. Soc.*, 1946, 109, 107-108.
8. Kerrich, J. E. The mathematical background. In Arbous, A. G. and Kerrich, J. E., *Accident statistics and the concept of accident proneness. National Institute for Personnel Research No. 391, 1951, South African Council for Scientific and Industrial Research.*
9. Maguire, R. A., Pearson, E. S., and Wynn, A. H. A. The time interval between industrial accidents. *Biometrika*, 1952, 39, 168-180.
10. Mintz, A. The inference of accident liability from the accident record. *J. appl. Psychol.*, 1954, 38, 41-46.
11. Mintz, A. and Blum, M. L. A reexamination of the accident proneness concept. *J. appl. Psychol.*, 1949, 33, 195-211.
12. Moran, P. A. P. The random division of an interval. *J. Roy. Stat. Soc.*, 1947, 9 (suppl.), 92-98.
13. Whitworth, W. A. *Choice and chance.* Reprint of 5th edit. Stechert, N. Y., 1934.

## Attitudes on Social Issues of Business Administrators and Students in a School of Business Administration

Alexis M. Anikeeff

Oklahoma A & M College

Students enrolled in a School of Business Administration generally plan to establish careers in business organizations, and aspire to become business administrators. Although considerable effort is expended in relating present student interests to future job success, the relationship of attitudes on significant social issues and future job success is apparently largely ignored. The general purpose of this study is to explore the possibility that student attitudes on social issues may serve as useful measures for prediction of success in various fields of endeavor. The specific purpose of this study is to measure the extent to which attitudes of business administrators differ from those of students in a School of Business Administration.

### Procedure

Members of a seminar distributed questionnaires containing 40 statements to 78 business administrators and to 146 business school students. Respondents were forced to reply either yes or no to each statement. Five statements dealt with unionism, 10 with government control, 15 with personnel policy, 5 with profit distribution, 4 with the free enterprise system, and 1 with the desirability of business training on the college level.

Business administrators were selected on the basis of the size of their establishments and their willingness to cooperate. The largest organization in the general vicinity of the seminar member's home was contacted first. When the top administrative officer was unavailable, the questionnaire was completed by an individual who was second in command. About 80% of the firms contacted employed less than 25 persons and all of the firms were located in Mississippi.

The student sample was composed entirely of Mississippi State College students who were enrolled in the School of Business and Industry. Approximately 60% of the students were upper-classmen. As in the case of business administrators, the student sample is composed entirely of students who were willing to cooperate with the survey.

### Results

Significant or very significant differences between responses of business administrators and students of business administration were found on 20 of the 40 statements contained in the questionnaire. Specific details are found in Table 1. The statements are numbered in accordance with their appearance in the questionnaire. The significance of the difference between percentages was estimated from the Lawshe and Baker nomograph.<sup>1</sup>

It is noteworthy that significant differences between responses of administrators and students were found on every government control statement which appeared in the questionnaire. In all cases the students professed to be more favorably disposed toward government regulation and control than the administrators.

### Discussion

Although a marked divergence of attitudes is indicated on one-half of the issues presented to both students and the administrators, the effect of this situation upon the future success of the students in their roles as administrators is not clearly discernible. Some shifting of attitudes may take place when students are placed on the job. There is also the possibility that some shifting of attitudes may occur on the part of administrators. In the event that student attitudes toward government control will not subject the students to unfavorable discrimination by present business administrators, it may be reasonable to believe that the forthcoming generation of administrators will be less prone to believe that the whole American economy will collapse with the further extension of governmental influence in business affairs.

<sup>1</sup> Lawshe, C. H. and Baker, P. C. Three aids in the evaluation of the significance between percentages. *Educ. psychol. Measmt.*, 1950, 10, 263-270.



Table 1  
Distribution of Responses to Items in Questionnaire

Item	Per Cent Replying Yes		Diff.
	Student	Adminis- trator	
1. Business should receive government subsidies.....	34%	19%	15%*
3. Corporations should be taxed higher than individuals.....	80	63	17**
4. Price control will destroy the free enterprise system.....	35	60	25**
7. Workers will do their best work only if strict discipline is maintained by the supervisor.....	30	46	16*
9. Employees should have company sponsored retirement plans.....	84	72	12*
11. More jobs should be covered by the minimum wage law.....	76	40	36**
13. Labor unions help industrial progress.....	75	60	15*
15. A worker who is "no good" on one job will probably be "no good" on any other job.....	10	24	14**
16. The federal government should subsidize educational institutions.....	58	40	18**
23. A Fair Employment Practices Commission should be established in Mississippi.....	55	33	18**
30. Shifty-eyed persons are dishonest.....	5	18	13**
31. Unemployment benefits should be abolished.....	13	36	23**
32. People with red hair are emotionally unstable.....	5	13	8*
33. You can tell a person's intelligence by interviewing him.....	26	42	16**
35. Government old age pensions should be abolished.....	13	23	10*
36. There should be absolutely no government control or regulation of pri- vate business.....	29	45	16*
37. Labor unions will destroy the free enterprise system.....	24	40	16**
38. Government should compete with private business whenever the public welfare can be improved.....	69	54	15*
39. Profits resulting from increased productivity should be divided equally among stockholders, labor, and the consumers.....	28	64	36**
40. You cannot have democracy without the free enterprise system.....	73	53	20**

\* Indicates 5% level of confidence.

\*\* Indicates 1% level of confidence.

### Summary

An attitude survey blank containing 40 statements and covering the areas of government control, personnel policy, profit distribution, labor unionism, and the free enterprise system was completed by 78 business administrators and 146 business administration students.

1. Significant differences were found on 20 of the 40 items contained in the questionnaire between responses of the two groups.

2. Disagreement was greatest in the area of government control. Students were significantly more favorably disposed to government control than the administrators.

3. Despite the marked divergence between attitudes of the two groups, the possibility exists that some student attitudes may shift toward those professed by administrators when the students are forced to solve problems presently faced by the administrators.

Received November 25, 1953.

## The Guilford Zimmerman Temperament Survey as a Predictor of Achievement Level and Achievement Fluctuation in Introductory Psychology

A. W. Bendig and J. L. Sprague

*University of Pittsburgh*

Student achievement in college courses where multiple measurements are taken over the semester involves two aspects: consistency and variability. Inter-student differences in the average level of course achievement over several examinations are used as the basis for assigning course grades and result in moderately reliable achievement indices (1). However, the reliability of average achievement levels is less than perfect due to intra-student fluctuation in performance from test to test. Glaser (4) has shown that measures of inconsistent responses in a retest situation have a moderate, but significant degree of reliability which suggests that measures of achievement fluctuation within a course may reliably measure an important aspect of student achievement performance.

We are suggesting that achievement fluctuation is particularly important in attempting to predict achievement level. The usual educational procedure of averaging scores or grades on several course tests to arrive at a final course grade implies that measures of achievement level and of achievement fluctuation will be related in a nonlinear manner. Students receiving A course grades are most likely to have shown consistent A or B performance on each test and those receiving F grades to have achieved at D or F levels on each test. However, the C or middle group includes both students who have received consistent C grades on each test and also those who have fluctuated widely from A to F on separate examinations, but whose average grade ends up as a C. We would predict a curvilinear relationship between measures of achievement level and achievement fluctuation, with the middle achievement level groups showing the largest average fluctuation and the extreme level groups (high and low) demonstrating significantly smaller fluctuation.

If this analysis is correct it then bears importantly on the problem of predicting student achievement in the first course in psychology. Aptitude test scores have shown moderate, but important rectilinear relationships with achievement level (2, 9, 10), but attempts to use personality and interest scales to predict level when aptitude is statistically held constant have been fruitless (7, 8). This lack of success may be due to: (a) curvilinear relationships existing between such scales and achievement level which are not revealed by rectilinear correlation techniques; or (b) these personality and interest scales being predictive of only the same type of student behavior that is predicted by aptitude tests. The first of these hypotheses could be tested by computing curvilinear correlation coefficients ( $\eta$  or  $\epsilon$ ) in addition to rectilinear Pearsonian coefficients and applying standard tests of significance to the difference between the pairs of rectilinear and curvilinear coefficients. Hypothesis (b) could be assessed by correlating the personality and/or interest scales with aptitude tests known to be related to the achievement criterion and finding the partial correlation of scales and achievement with aptitude statistically held constant. However, our assumed relationship between achievement level and fluctuation suggests a third hypothesis: (c) a predictor may be rectilinearly related to *both* level and fluctuations, but because of the curvilinear confounding of level and fluctuation may show a zero correlation with level. This third hypothesis could be tested by correlating each scale with measures of both level and fluctuation and tempering our judgments of nonsignificant scale-level correlations in the light of obtained scale-fluctuation relationships.

Several recent studies have suggested that the Guilford Zimmerman Temperament Sur-

vey (5) may be useful in predicting student achievement in introductory psychology. Klugh (6) found three scales on the GZTS to correlate positively and significantly with total scores on the ACE. For a sample of 225 male students the Objectivity and Friendliness scales were significant at the .01 level ( $r = .19$  and  $.18$ ), while the Personal Relations scale was significant at the .05 level ( $r = .14$ ). Of the remaining seven scales, only the Masculinity scale approached significance ( $r = .11$ ). Since the ACE is significantly related ( $R = .47$ ) to achievement level in introductory psychology (10) these GZTS scales could be expected to show some correlation with the same achievement variable. However, Krumm (8), using the ACE as a predictor, obtained discrepancy scores between the predicted and obtained grades in introductory psychology ( $N = 410$ ) and identified the top and bottom quarters of the resulting distribution. Comparing the mean GZTS scores of these groups of "over-achievers" and "underachievers" showed none of the GZTS scales to significantly discriminate between these extreme groups.

The problem of the present research was to compare the relation of GZTS to achievement level in introductory psychology when both rectilinear and curvilinear correlation techniques are used, and to make a similar comparison when achievement fluctuation is used as the criterion.

### Procedure

*Subjects.* The Ss were 155 students enrolled in daytime sections of introductory psychology at the University of Pittsburgh during the Spring, 1953, semester and who were present in class the day the personality questionnaire was administered. There were 107 men and 48 women in the sample with the great majority being freshmen and sophomores.

*Variables.* The Guilford Zimmerman Temperament Survey (5) was administered to all sections near the beginning of the semester by trained examiners.<sup>1</sup> Raw scores on each of the ten GZTS scales were used in the later analysis.

The achievement variables were derived from students' scores on five course achievement examinations given during the semester. All five

tests were objective, 50-item, multiple-choice examinations and the raw scores from each test were converted to standard scores based upon the performance of all students in introductory psychology on each single test. The large majority of students ( $N = 126$ ) took all five tests during the semester, but graduating seniors ( $N = 22$ ) were excused from the last two tests. The remaining students ( $N = 7$ ) had missed one of the tests and had received an incomplete grade in the course, but were retained in the present sample. Details of the course evaluation procedure have been previously described (1).

The variable of achievement level used in this study was the letter grade received by each student. These grades were determined by finding the mean of each S's test standard scores and converting this average to a letter grade on the basis of previously established cutting points which were common to all sections. This achievement level variable has been shown to have an estimated reliability of .80 (1, p. 316).

The achievement fluctuation variable was derived from the range between the highest and lowest test scores received by each S over the semester. Dixon and Massey (3, pp. 240-241) have shown that the range in small samples is a highly efficient estimate of the population variability and is computationally much simpler than computing the standard deviation of each S's scores. The reliability of this fluctuation measure was estimated by drawing a random sample of 100 Ss who had taken all five course tests. The range was computed for each S as the difference between his highest and lowest scores and a second measure of the fluctuation was obtained by finding the range between the S's second highest and second lowest test scores. Correlating these pairs of fluctuation measures for the 100 Ss gave a coefficient of .59, which indicates that the range is a relatively stable measure of achievement fluctuation. Finally, the range for each of the 155 Ss in the study sample was multiplied by the constants given by Dixon and Massey (3, p. 240) to give an estimate of the standard deviation of each S's achievement test scores.

### Results

The distribution of achievement fluctuation measures appeared positively skewed and was tested for normality by grouping individual measures into six categories and applying the usual chi-square test. Chi-square equalled 9.94 which with 3 degrees of freedom was significant at the .05 level. To reduce skewness these measures were put through a square-root transformation and the transformed distribution tested for normality by the usual chi-square method. The chi-

<sup>1</sup> Appreciation is expressed to Dr. Frederick Herzberg of Psychological Services of Pittsburgh who supervised the administration and scoring of the temperament survey.



square value with three degrees of freedom was 5.60 which is not significant at the .05 level of confidence. The mean and variance of the fluctuation measures were computed for each of the five achievement level groups and an analysis of variance performed to test the hypothesized curvilinear relationship between level and fluctuation. The F comparison between the means gave an F value of 4.03 which is significant at the .01 level. This F corresponds to a curvilinear correlation ( $\eta^2$ ) of .31, while the product-moment rectilinear correlation between level and fluctuation gave a nonsignificant  $r$  of  $-.15$ . A chi-square test of the significance of the difference between their curvilinear and rectilinear coefficients gave a chi-square of 12.40, which, with 3 degrees of freedom, is significant at the .01 level. A chi-square test of the homogeneity of the variances of the fluctuation measures within the five achievement level groups yielded a value of 5.61 which, with four degrees of freedom, is not significant at the .05 level. The mean fluctuation for each of the achievement level groups can be found in Table 1. As hypothesized, the middle achievement level groups show the largest average fluctuation and the extreme level groups (A and F) demonstrated significantly less average fluctuation.

Since our two measures of achievement, level and fluctuation, are nonlinearly related,

it was necessary to perform a further transformation on the fluctuation measures to clarify the relation of GZTS scales to these two criteria. To insure a zero correlation between level and fluctuation each S's fluctuation measure was taken as a deviation (plus or minus) from the mean fluctuation of his achievement level group. Since the variances of the fluctuation measures within the level groups appeared to be homogeneous, the further step of dividing each deviation by the standard deviation of its level group was not necessary. These deviations from all five level groups were then pooled and divided into five fluctuation groups. Group I included the 20 Ss showing the largest intra-subject achievement fluctuation (independent of achievement level), Group V comprised 17 Ss with the smallest intra-subject fluctuation, with Groups II, III, and IV consisting of Ss showing intermediate amounts of fluctuation.

Raw scores on the ten GZTS scales were then correlated with the criterion measures of achievement level and achievement fluctuation. Rectilinear product-moment correlations were computed by weighting achievement level groups A through F with unit digits 4 through 0 and similarly weighting achievement fluctuation groups I through V with the same weights. These weights were then correlated with the raw GZTS scores. In addition, curvilinear correlations ( $\eta^2$ ) between the GZTS scales and the two criterion measures were computed and chi-square tests of the significance of the difference between the rectilinear and curvilinear coefficients evaluated. These correlations and tests of curvilinearity are given in Table 2. It can be noted that the GZTS Objectivity scale is rectilinearity related to achievement level and this is probably also true of the Restraint scale. Friendliness and Masculinity are related to level in a curvilinear fashion, but the product-moment coefficients for these two scales are not significant. None of the GZTS scales are related to achievement fluctuation when only the product-moment coefficients are considered, but Ascendancy, Social Interest, and Emotional Stability are curvilinearly related to fluctuation.

Table 1

Distribution of Achievement Fluctuation Measures  
for the Achievement Level Groups

Achievement Level	Number of Subjects	Fluctuation Mean	Standard Deviation
A	21	3.12	.91
B	32	3.68	.91
C	62	3.83	.88
D	26	4.05	.60
F	14	3.49	.92
Significance of Means (F)		4.03**	
Homogeneity of Variances (chi-square)			5.61

\*\* Significant at the .01 level.

Table 2  
Rectilinear and Curvilinear Correlations between Guilford-Zimmerman Scales and Achievement Level and Fluctuation

GZTS Scale	Achievement Level			Achievement Fluctuation		
	Product-Moment Correlation ( <i>r</i> )	Curvilinear Correlation (Eta)	Significance of Difference (Chi-Square)	Product-Moment Correlation ( <i>r</i> )	Curvilinear Correlation (Eta)	Significance of Difference (Chi-Square)
General Activity	-.13	.18	2.55	-.02	.18	4.83
Restraint	.20*	.24*	2.56	-.03	.08	0.93
Ascendancy	-.13	.19	2.80	.00	.35**	18.08**
Social Interest	-.14	.20	2.98	.11	.27*	9.20*
Emotional Stability	.13	.17	2.13	.05	.24*	7.98*
Objectivity	.21*	.28*	5.34	.03	.07	0.58
Friendliness	.11	.27*	9.16*	.15	.17	1.29
Thoughtfulness	-.02	.04	0.18	.09	.14	1.84
Personal Relations	.08	.19	4.68	.12	.17	2.46
Masculinity	.15	.25*	5.79	-.09	.12	0.90

\* Significant at the .05 level.

\*\* Significant at the .01 level.

### Discussion

Our results confirm the hypothesis of a significant curvilinear relation between achievement level and achievement fluctuation. This indicates that our level criterion is an impure measure of achievement differences between students and suggests that similar criteria used widely in educational research are similarly contaminated by the fluctuation variable. Nor is intra-student fluctuation a chance phenomenon: the moderate but significant correlation between two similar measures of fluctuation that was found in this study shows its reliability.

The correlations in Table 2 bear on points (a) and (c) made in the third paragraph of this paper. Two GZTS scales, Restraint and Objectivity, are rectilinearly related to our contaminated level criterion, but two additional scales, Friendliness and Masculinity, show insignificant rectilinear, but significant curvilinear correlations with level. This confirms point (a), since neither of these last two scales would have appeared to be related to level if curvilinear correlation techniques had not been used. However, point (c), as expressed in the second paragraph, is not confirmed, since none of the GZTS scales are rectilinearly related to fluctuation. The

significant curvilinear correlation of three GZTS scales, Ascendancy, Social Interest, and Emotional Stability, with fluctuation suggests that point (c) is too naively stated. Omitting the word "rectilinearly" in point (c) yields a hypothesis that appears plausible in view of our findings. Perhaps these three GZTS scales show essentially zero relationships with the impure criterion of level because of their curvilinear correlation with the pure measure of fluctuation. These results do not confirm point (c), but suggest that a modified form of the hypothesis is tenable.

The available data did not permit a direct test of point (b). However, there are suggestive consistencies and discrepancies between our results and previous studies (6, 8, 10) that indirectly bear on this point. Klugh (6) found the Objectivity, Friendliness, and Personal Relations scales to be significantly related to the total score on the ACE, and the Masculinity scale to fall just short of statistical significance. Russell and Bendig (10) demonstrated a significant relation between the ACE and our level criterion, while Krumm (8) showed the GZTS scales were not predictive of achievement level in introductory psychology when the variability in level attributable to ACE differences was statistically elimi-

nated. We found the Restraint, Objectivity, Friendliness, and Masculinity scales significantly related to level when academic aptitude is uncontrolled. These results suggest that the Friendliness and Objectivity scales on the GZTS measure the same aspects of achievement performance that is measured by the ACE and could not profitably be used in a regression equation along with the ACE to predict achievement level in introductory psychology. However, the Restraint and Personal Relations scales probably would increase the predictability of level if used in conjunction with the ACE: the Restraint scale because of its significant correlation with level and its low correlation with ACE, while the Personal Relations scale could act as a suppressor variable due to its lack of correlation with level and its significant relation to the ACE. Admittedly the predictive usefulness of these two scales is unproven, but provides a hypothesis to be tested in a later sample.

#### Summary

Scores on the Guilford-Zimmerman Temperament Survey were correlated by both rectilinear and curvilinear methods with measures of course achievement level and intra-student achievement fluctuation in introductory psychology ( $N = 155$ ). Achievement level and fluctuation were curvilinearly related and the fluctuation measures were adjusted to remove this artifact. Two GZTS scales, Restraint and Objectivity, were rectilinearly related to level ( $r = .20$  and  $.21$ ), while two additional scales, Friendliness and Masculinity, showed significant curvilinear correlations with level ( $\eta = .27$  and  $.25$ ).

None of the GZTS scales were rectilinearly related to fluctuation, but three scales, Ascendancy, Social Interest, and Emotional Stability, were curvilinearly correlated with fluctuation ( $\eta = .35$ ,  $.27$ , and  $.24$ ).

Received November 27, 1953.

#### References

1. Bendig, A. W. The reliability of letter grades. *Educ. psychol. Measmt.*, 1953, 13, 311-321.
2. Carlson, H. B., Fischer, R. P., and Young, P. T. Improvement in elementary psychology as related to intelligence. *Psychol. Bull.*, 1945, 42, 27-34.
3. Dixon, W. J. and Massey, F. J. *Introduction to statistical analysis*. New York: McGraw-Hill, 1951.
4. Glaser, R. The reliability of inconsistency. *Educ. psychol. Measmt.*, 1952, 12, 60-64.
5. Guilford, J. P. and Zimmerman, W. S. *Guilford-Zimmerman Temperament Survey Manual of Instructions and Interpretations*. Beverly Hills, California: Sheridan Supply Co., 1949.
6. Klugh, H. E. The relationship between some aspects of temperament and academic aptitude. (Unpublished study.)
7. Klugh, H. E. The prediction of academic achievement from measures of personality. Unpublished master's thesis, Univer. of Pittsburgh, 1952.
8. Krumm, R. L. Interrelationships of measured interests and personality traits of introductory psychology instructors and their students as related to student achievement. Unpublished doctor's dissertation, Univer. of Pittsburgh, 1952.
9. Newman, S. E., Duncan, C. P., Bell, G. B., and Bradt, K. H. Predicting student performance in the first course in psychology. *J. educ. Psychol.*, 1952, 43, 243-247.
10. Russell, H. E. and Bendig, A. W. Student ratings of instructors and course achievement with academic aptitude controlled. *Educ. psychol. Measmt.*, 1953, 13, 626-635.



## Proposed Hostility and Pharisaic-Virtue Scales for the MMPI \*

Walter W. Cook

*University of Minnesota*

and

Donald M. Medley

*Indiana University*

This article describes an attempt to develop scales for the *Minnesota Multiphasic Personality Inventory* (MMPI) which measure a person's ability to get along well with others. Such scales should prove valuable in selecting personnel who must deal with the public or work harmoniously and effectively with a group. The validity of the scales reported here is based on their power to predict the rapport of teachers with pupils in a classroom. Since the content of the items is not directly related to school work, it is believed that the scales may prove useful also in the selection of sales people, officers and non-commissioned officers in the armed forces, foremen, and other personnel who must be able to establish rapport with others and maintain group morale. The scales are published here with a view to encouraging further experimentation in other situations.

For a number of years a series of research studies, centering at the University of Minnesota, has been carried on with the isolation and measurement of non-intellectual factors related to success in teaching as its principal object. A major outcome of this work has been the development of the *Minnesota Teacher Attitude Inventory* (MTAI) (2), which has been found to predict teacher-pupil rapport with a degree of validity indicated by correlations with independent criteria of from .50 to .63. When the MTAI was standardized on a large sample of Minnesota teachers (1), it was possible to identify, in the extremes of the distribution, two groups of teachers sharply differing in their ability to get along with pupils. The MMPI (3) was administered to these two groups, and

212 completed inventories, 112 representing approximately the 8 per cent of teachers scoring highest, and 100 the 8 per cent scoring lowest (among all of the public school teachers in Minnesota) on the MTAI obtained.

The MMPI contains 550 items of the True-False type with a wide variety of content. When the proportions making each response in each of the two groups of teachers were compared, after being transformed to angles by the arc-sine transformation (5), the difference between the groups was found to be significant at the 5 per cent level on 250 items.

The teacher who scores low on the MTAI describes himself in his responses (2) as generally hostile toward others; he says that pupils are dishonest, insincere, untrustworthy, lazy, etc. His self-description also indicates that he: (a) adheres excessively to rigid standards of morality; (b) tends to dominate those below him and be subservient to those above him; and (c) prides himself on a thorough knowledge of his subject-matter. Among the 250 discriminating MMPI items were many which reflected generalized hostility toward people, and others that suggested *Pharisaic virtue*. There were no items which reflected the tendency toward security through power over people or through mastery of subject matter. The other discriminating items suggested symptoms of depression, anxiety and general neurosis.

A total of 77 items which most obviously reflected hostility were chosen for a preliminary "Ho" scale, and 60 items having to do with virtue and morality were chosen for a "Pv" scale. When the MMPI answer sheets completed by 200 graduate students in education (all of whom were experienced classroom teachers) were scored on these two

\* This study was made possible by a grant from the research funds of the Graduate School of the University of Minnesota.

keys, correlations of  $-.45$  (for "Ho") and  $-.49$  (for "Pv") with the MTAI were obtained. On the strength of these results, further refinement of the scales was undertaken. Five clinical psychologists, working independently, selected sets of "Ho" and "Pv" items. On the basis of agreement among the five, a final 50-item Ho key was selected. The reliability coefficient of this scale for the 200 graduate students, estimated by analysis of variance (4), was  $.86$ .

Substantial agreement among the psychologists could be obtained on only 20 of the Pv items. On the basis of an internal consistency item analysis carried out on the 200 graduate students, 30 items were added to produce a 50-item Pv key. The internal con-

Table 1

Relationships Among Ho, Pv, and MTAI Scales

Correlation Coefficients	Males N = 100	Females N = 100	Total N = 200
Ho vs. MTAI	$-.44$	$-.45$	$-.44$
Pv vs. MTAI	$-.38$	$-.54$	$-.46$
Ho vs. Pv	$.65$	$.73$	$.69$
Ta vs. MTAI	$-.45$	$-.54$	$-.50$
Multiple R, Pv + Ho vs. MTAI Regression coefficients	$-.46$	$-.55$	$-.51$
Beta weight for Ho	$-.335$	$-.109$	
Beta weight for Pv	$-.163$	$-.463$	

sistency of this scale could not be estimated on the 200 papers used in the item analysis, so the papers of 55 other graduate students in education were used for this purpose. The reliability, estimated by analysis of variance, was  $.88$ .

Direct evidence regarding the validity of the Ho and Pv scales for predicting pupil-teacher rapport as measured by the MTAI, and indirect evidence as to their validity for measuring "Hostility" and "Pharisaic virtue," was obtained by correlating the scores of the 200 graduate students on the two scales with their scores on the MTAI. The results are summarized in Table 1.

The sample contained 100 males and 100 females; correlations and beta weights are presented for the two sexes separately in the

Table 2

Items Included in the Ho (Hostility) Key for the Minnesota Multiphasic Personality Inventory \*  
(Listed according to number on the Group Form)

19	136	265	386	455
28	148	271	394	458
52	157	278	399*	469
59	183	280	406	485
71	226	284	410	504
89	237*	292	411	507
93	244	319	426	520
110	250	348	436	531
117	252	368	438	551
124	253*	383	447	558

\* Items marked with an asterisk are keyed "False"; all other items are keyed "True."

first two columns, and correlations for the entire sample in the third column.

The Ho scale tends to be more effective for males than the Pv scale, while the reverse holds for females, although none of the sex differences is statistically significant. In the multiple regression equation for predicting MTAI scores from Ho and Pv scores for males, the addition of the Pv scale does not significantly improve on the prediction from the Ho scale alone. In the multiple regression equation for predicting MTAI scores from Ho and Pv scores for females, the addition of the Ho scale does not significantly improve on the prediction from the Pv key alone.

Table 3

Items Included in the Pv (Pharisaic Virtue) Key for the Minnesota Multiphasic Personality Inventory \*  
(Listed according to number on the Group Form)

13	147	356	401*	468
26	158	357	402	470
30*	176*	361	404	492
45*	206	375	413	499
58	232	378	414	502
94	289	380	416	506
111	317	390	439	509
112	336	392	443	510
119	337	395	457	548
129	338	397	461	564

\* Items marked with an asterisk are keyed "False"; all other items are keyed "True."

Table 4  
Norms for Hostility (Ho) Scale of MMPI

Raw Score	T Score		Raw Score	T Score		Raw Score	T Score	
	M	F		M	F		M	F
50	93	95	33	69	71	16	46	47
49	91	94	32	68	70	15	44	46
48	90	92	31	66	68	14	43	45
47	89	91	30	65	67	13	41	43
46	87	89	29	64	66	12	40	42
45	86	88	28	62	64	11	39	40
44	84	87	27	61	63	10	37	39
43	83	85	26	59	61	9	36	38
42	82	84	25	58	60	8	35	36
41	80	82	24	57	59	7	33	35
40	79	81	23	55	57	6	32	33
39	77	80	22	54	56	5	30	32
38	76	78	21	53	54	4	29	31
37	75	77	20	51	53	3	28	29
36	73	75	19	50	52	2	26	28
35	72	74	18	48	50	1	25	26
34	71	73	17	47	49	0	23	25

The correlations obtained with multiple regression weights on both scales combined are practically identical with those obtained when the two scales are thrown together into one 100-item "Ta" (teacher attitude) scale. If

the best predictor of teacher-pupil rapport for both sexes is desired, the Ta scale would probably be the most satisfactory.

The magnitude of the intercorrelation between the two scales is enough smaller than

Table 5  
Norms for Pharisaic-Virtue (Pv) Scale of MMPI

Raw Score	T Score		Raw Score	T Score		Raw Score	T Score	
	M	F		M	F		M	F
50	99	91	33	73	66	16	46	41
49	98	90	32	71	64	15	45	39
48	96	88	31	70	63	14	43	38
47	94	87	30	68	62	13	41	36
46	93	85	29	66	60	12	40	35
45	91	84	28	65	59	11	38	34
44	90	82	27	63	57	10	37	32
43	88	81	26	62	56	9	35	31
42	87	79	25	60	54	8	34	29
41	85	78	24	59	53	7	32	28
40	84	76	23	57	51	6	31	26
39	82	75	22	55	50	5	29	25
38	80	73	21	54	48	4	27	23
37	79	72	20	52	47	3	26	22
36	77	70	19	51	45	2	24	20
35	76	69	18	49	44	1	23	19
34	74	67	17	48	42	0	21	17



their reliabilities to suggest that they are measuring different, although highly related, dimensions of personality.

Lists of the items included in the two scales are presented as Tables 2 and 3, all items being keyed "true" except those marked with an asterisk. The numbers given are those on the group form of the MMPI (3). A key for either of the scales may be easily prepared by making a scoring stencil perforated as indicated in these tables.

If it is remembered that these items represent the individual's own description of him-

self, some insight into the personality of the individual who scores high on one of these scales may be obtained by reading the items.

Typical items on the Ho scale are the following: "I would certainly enjoy beating a crook at his own game," "When someone does me a wrong I feel I should pay him back if I can, just for the principle of the thing," "I have often met people who were supposed to be expert who were no better than I." Thus revealed, the hostile person is one who has little confidence in his fellowman. He sees people as dishonest, unsocial, immoral, ugly,

Table 6

Norms for the Teacher Attitude (Ta) Scale (Ho plus Pv Scales) of MMPI

Raw Score	T Score		Raw Score	T Score		Raw Score	T Score	
	M	F		M	F		M	F
100	100	98	66	73	71	33	46	44
99	99	97	65	72	70	32	45	44
98	99	96	64	71	69	31	45	43
97	98	95	63	70	68	30	44	42
96	97	95	62	70	67	29	43	41
95	96	94	61	69	67	28	42	40
94	95	93	60	68	66	27	41	40
93	95	92	59	67	65	26	41	39
92	94	91	58	66	64	25	40	38
91	93	91	57	66	64	24	39	37
90	92	90	56	65	63	23	38	36
89	91	89	55	64	62	22	37	36
88	91	88	54	63	61	21	37	35
87	90	87	53	62	60	20	36	34
86	89	87	52	62	60	19	35	33
85	88	86	51	61	59	18	34	33
84	87	85	50	60	58	17	33	32
83	86	84	49	59	57	16	33	31
82	86	83	48	58	56	15	32	30
81	85	83	47	58	56	14	31	29
80	84	82	46	57	55	13	30	29
79	83	81	45	56	54	12	29	28
78	82	80	44	55	53	11	29	27
77	82	79	43	54	52	10	28	26
76	81	79	42	53	52	9	27	25
75	80	78	41	53	51	8	26	25
74	79	77	40	52	50	7	25	24
73	78	76	39	51	49	6	24	23
72	78	75	38	50	48	5	24	22
71	77	75	37	49	48	4	23	21
70	76	74	36	49	47	3	22	21
69	75	73	35	48	46	2	21	20
68	74	72	34	47	45	1	20	19
67	74	71						

and mean, and believes they should be made to suffer for their sins. Hostility amounts to chronic hate and anger.

Among the 20 "core" items on the Pv scale are items like the following: "I believe that a person should never taste an alcoholic drink," "Sexual things disgust me," and "I deserve severe punishment for my sins," suggesting preoccupation with ideas of sin and punishment; among the 30 items added by item analysis such items as "I am inclined to take things hard," "It makes me nervous to have to wait," and "Dirt frightens or disgusts me," suggest general neurosis.

Norms for the Ho, Pv, and Ta scales were derived on a sample of the same normal group that was used in deriving the norms for the original clinical scales of the MMPI. The sample consisted of 541 individuals, 226 males and 315 females. These norms for males and females are presented in Tables 4, 5, and 6.

### Summary

The development of two keys for the *Minnesota Multiphasic Personality Inventory* by selecting principally on the basis of content two sets of 50 items from 250 found to discriminate significantly between teachers scoring high and teachers scoring low on the *Minnesota Teacher Attitude Inventory* is described and the items are listed. The Ho scale (Hostility) reveals a type of individual

characterized by a dislike for and distrust of others. The Pv scale (Pharisaic virtue) reveals a type of person who describes himself as preoccupied with morality and ridden with fears and tensions. A Ta (Teacher attitude) scale made up of all 100 items is also proposed. When administered to a rather homogeneous group of graduate students in education classes, the internal consistency reliability coefficients of the two short scales were estimated to be .86 (for Ho) and .88 (for Pv), and the Ho, Pv, and Ta scales correlated — .44, — .46, and — .50, respectively, with the *Minnesota Teacher Attitude Inventory*.

Received November 3, 1953.

### References

1. Cook, W. W. and Hoyt, C. J. Procedure for determining number and nature of norm groups for the *Minnesota Teacher Attitude Inventory*. *Educ. psychol. Measmt.*, 1950, 12, 562-573.
2. Cook, W. W., Leeds, C. H., and Callis, R. *Minnesota Teacher Attitude Inventory*. New York: The Psychological Corporation, 1951.
3. Hathaway, S. H. and McKinley, J. C. *The Minnesota Multiphasic Personality Inventory*. Minneapolis, Minnesota: The University of Minnesota Press, 1943.
4. Hoyt, C. J. Test reliability obtained by analysis of variance. *Psychometrika*, 1941, 6, 153-160.
5. Zubin, J. A transformation function for proportions and percentages. *J. appl. Psychol.*, 1935, 19, 213-220.

## The Relationship of Job Values and Desires to Vocational Aspirations of Adolescents

Stanley L. Singer

*Valley Psychological Consultants, Van Nuys*

and Buford Steffire

*Counseling and Guidance Service, Los Angeles Board of Education*

Two relatively recent developments in counseling theory relate to the importance of understanding the individual's "level of vocational aspiration" and "job values and desires." These two personality dimensions are assuming increasing importance in our attempts to explore the dynamics of vocational selection and adjustment.

Vocational counselors have long been aware of the importance of level of vocational aspiration as a guidepost in making long range plans because realism in level of aspiration does much to overcome pressures for the selection of vocational goals which might lead to much frustration. Because of the importance of vocational aspiration level to mental health, research in this area is greatly needed, particularly as an aid in disclosing the relationship of aspiration level to other aspects of vocational selection.

Another area where research is needed is that of job values and desires which are also of great importance in making vocational plans. By job values and desires are meant the answers given to the basic question, "What do you really want from a job?" Job values and desires refer not to the kind of work or duties performed, but to the source of satisfaction in the work and are defined in this study as the choices listed in the following *Job Values and Desires Checklist*.

### Centers' Job Values and Desires Checklist

If you had a choice of one of these kinds of jobs, which would you choose? (Put a number "1" by your FIRST choice. If you have OTHER choices which you would like to indicate, put a number "2" by your second choice and a number "3" by your third.)

- A. A job where you could be a leader.
- B. A very interesting job.

- C. A job where you would be looked upon very highly by your fellow men.
- D. A job where you could be boss.
- E. A job which you were absolutely sure of keeping.
- F. A job where you could express your feelings, ideas, talent, or skill.
- G. A very highly paid job.
- H. A job where you could make a name for yourself—or become famous.
- I. A job where you could help other people.
- J. A job where you could work more or less on your own.

Centers<sup>1</sup> has done extensive work on this problem with adults from different social classes as well as from rural and urban environments. His major finding was that self-expression is a "middle class" job value while security is a "working class" value.

The present study attempts to examine the job values and desires of adolescents in relation to level of aspiration as measured by the Level of Interest section of the California Occupational Interest Inventory. The problem being explored here is whether differences in level of aspiration are reflected in differences in job values and desires. It is hoped that some understanding of the relationship between the concepts of level of interest and of job values may follow from such an exploration.

Some justification is needed for considering the Level of Interest section as a measure of vocational aspiration. The manual<sup>2</sup> for the interest inventory gives no evidence indicating that scores on the Level section are in any way associated with differences in as-

<sup>1</sup> R. Centers. *Psychology of social class*. Princeton, New Jersey: Princeton University Press, 1949, 219 pp.

<sup>2</sup> E. A. Lee and L. A. Thorpe. *Manual of directions—Occupational Interest Inventory*, Advanced Series. Hollywood: California Test Bureau, 1943.



piration level. However, Steffire<sup>3</sup> found that when scores on the Level of Interest section were compared to an independent measure of vocational aspiration—the client's vocational objective—subjects aspiring to the higher level occupations made significantly higher scores. Steffire concluded, in speaking of the Level section, "This section of the test would appear to be a good rough index of the direction and extent of the student's aspiration as it will be expressed through the selection of a vocational objective." This research on over 1,000 high school seniors suggests that the Level of Interest section on the Lee-Thorpe Occupational Interest Inventory is an adequate measure of vocational aspiration.

The present study compared the job values and desires of seventeen- and eighteen-year-old Caucasians who scored in the lower quarter on the Level of Interest section of the California Occupational Interest Inventory to similar groups scoring in the upper quarter on the same section. The null hypothesis is that differences in level of aspiration are unrelated to the preference for job values and desires. The sample was composed of 212 male high school seniors and 242 female high school seniors from the Los Angeles City Schools. Separate analyses were made for males, for females, and for a combined sample of both sexes. Chi square with the Yates correction was applied to examine the relationships.

All subjects had participated in a specialized vocational guidance program made available to them during the 1952-53 school year. The guidance program consisted of seven steps: (1) initial structuring meeting during which the entire counseling program was explained; (2) basic testing which measured mental capacity, interest, and temperament; (3) initial interview with a counselor to relate test results and personal-social background to tentative vocational objectives; (4) study of occupational information; (5) additional testing as needed; (6) final interview to plan objectives and training; and

(7) invitation to the parents to discuss the student's plans with the counselor.

During the basic testing period, the measure of interest used was the California Occupational Interest Inventory. This test has six fields of interest: Personal-Social, Natural, Mechanical, Business, Arts, and Science—three types of interest—Verbal, Manipulative, and Computational—and a Level of Interest section which has been discussed above as a measure of vocational aspiration. The present study was only concerned with this last section of the test.

Centers' *Job Values and Desires Checklist* was used as the index of the student's job value preferences. The card was presented and checked during the first interview.

Consistently the percentage of respondents selecting category B—"Interesting experience"—and category F—"Self-expression"—was far above the percentage selecting any of the other categories. This finding was apparent for both the males and females as well as for the combined group. Only the lower quarter male group did not show a strong preference for "self-expression." The three categories selected least often were "power," "leadership," and "esteem."

Table 1 summarizes the results for the males. Chi square was significant in two of

Table 1  
Chi Square of Upper and Lower Quarters on Level of Interest Section and Job Values and Desires for Males

Category	Upper Quarter (N = 148) %	Lower Quarter (N = 64) %
A. Leadership	5	2
B. Interesting Experience	18	27
C. Esteem	2	3
D. Power	4	5
E. Security	12	12
F. Self-Expression	29	9**
G. Profit	11	8
H. Fame	4	6
I. Social Service	7	9
J. Independence	8	19*

\* Significant at 5 per cent level.

\*\* Significant at 1 per cent level.

<sup>3</sup> B. Steffire. *Psychological factors associated with aspiration for socio-economic mobility*. Unpublished dissertation, University of Southern California, June 1953.

Table 2

Chi Square of Upper and Lower Quarters on Level of Interest Section and Job Values and Desires of Females

Category	Upper Quarter (N = 137) %	Lower Quarter (N = 105) %
A. Leadership	4	0
B. Interesting Experience	26	31
C. Esteem	2	4
D. Power	1	1
E. Security	6	11
F. Self-Expression	28	19
G. Profit	4	4
H. Fame	4	6
I. Social Service	18	17
J. Independence	7	7

Note: No differences between upper and lower quarters were significant.

the comparisons. On category F—"A job where you could express your feelings, ideas, talent, or skill"—P was significant beyond the 1 per cent level of confidence. Selection of this value is positively related to high vocational aspiration level. On item J—"A job where you could work more or less on your own"—chi square was significant at the 5 per cent level of confidence. Here the males falling in the bottom quarter on vocational aspiration tended to select the value of job "independence" more often than the group in the upper quarter.

Table 2 summarizes the findings for the females. It is apparent from the results that scores on the Level of Interest section had no significant relationship to the selection of job values and desires.

Table 3 presents the results for the combined group of males and females. Two of the comparisons were statistically significant. Category A—"A job where you could be a leader"—was preferred by more subjects than would be expected who fell in the upper quarter on the aspiration measure. By the same token, those falling in the lower quarter tended to underselect this particular value. "Leadership," it will be recalled, showed no relationship to vocational aspiration score when considered for males and females sepa-

ately, and was one of the categories least selected by all groups.

Category F—"A job where you could express your feelings, ideas, talent, or skill"—was significantly overselected by the group scoring in the upper quarter on the Level section while this same job value was of little concern to those scoring in the bottom quarter on the aspiration measure. It will be recalled that "self-expression" was significantly related to score on Level of Interest for the males also, although not for the females.

Summarizing the findings then, males who demonstrate high level of vocational aspiration are relatively more concerned with job values and desires that involve "self-expression." On the other hand, males who demonstrate low vocational aspiration are relatively more concerned with the job value of "independence." For adolescent females there appears to be no significant relationship between aspiration level and job values. For the combined group of males and females, desires for "leadership" and "self-expression" are positively related to high vocational aspiration.

The negative findings for females may mean that adolescent girls select job values from very personal motives unrelated to aspirations for social status. Since the eventual

Table 3

Chi Square of Upper and Lower Quarters on Level of Interest Section and Job Values and Desires for Combined Group

Category	Upper Quarter (N = 285) %	Lower Quarter (N = 169) %
A. Leadership	5	1*
B. Interesting Experience	22	30
C. Esteem	2	4
D. Power	2	2
E. Security	9	12
F. Self-Expression	28	15**
G. Profit	8	5
H. Fame	4	6
I. Social Service	12	14
J. Independence	8	11

\* Significant at 5 per cent level.

\*\* Significant at 1 per cent level.

socio-economic status of a girl is more likely to be determined by marriage than by her occupation, it is possible that her strivings for social status are not reflected in vocational values and desires.

In a review of the findings, certain similarities to Centers' results become apparent. It must be kept in mind in making this comparison that the present study examined the relation of expressed job values to a Level of Interest scale whose connection with ultimate occupational status and socio-economic status has not been established, while Centers studied the relation of expressed job values to known socio-economic status (middle class or working class status). The present study found a preference for "self-expression" (in males and in combined sex sample) and for "leadership" (in combined sex sample) to be related to a high level of vocational aspiration; Centers found a preference for "self-expression" and to some extent, "leadership" to be related to membership in the middle class. The present study found preference for "independence" to be related to a low level of vocational aspiration; Centers noted a tendency for "leadership" preference to be related to membership in the working class. The findings of the two studies, when compared in this manner, suggest the need to ex-

amine more closely the relationship between level of vocational interest in adolescents and socio-economic status. It is possible that the adolescent with a high level of vocational aspiration identifies himself with the middle class and hence views job values in the manner of adult middle class members while the adolescent with a low level of vocational aspiration may identify himself with the working class.

### Summary

This study has examined the relationship between: (1) level of aspiration, as measured by the Level of Interest section of the Occupational Interest Inventory; and (2) job values and desires, as measured by the checklist developed by Centers. For the male group it was demonstrated that a relationship exists between these two variables for some job values and that this relationship seems to be in line with that noted by Centers when he examined the role of socio-economic differences in the selection of job values and desires. Such a finding gives some tentative and indirect support to the belief that scores on the Level of Interest section may indicate the socio-economic status with which the adolescent male identifies himself.

*Received December 3, 1953.*



## Permanence of Strong Vocational Interest Blank Scores<sup>1</sup>

Kalmer E. Stordahl

University of Minnesota<sup>2</sup>

In assisting young men and women to make appropriate vocational and educational choices counselors make extensive use of tested interests. One of the problems with which both the counselor and counsellee are concerned is the permanence or stability of the scores on the measuring instrument.

The present study was designed to give an estimate of the permanence of the scores of pre-college males on Strong's Vocational Interest Blank. For an excellent review of the literature up to 1943 on the permanence of Strong scores see Strong (7). Two recent studies of the permanence of scores over a number of years have also been published by Strong (5, 6).

### Method

In the spring of 1949, the Vocational Interest Blank was offered on an optional basis to all high school seniors who participated in the state-wide testing program in the state of Minnesota. Approximately 3,500 senior boys completed the blank.

A check was made of the University enrollment in the spring of 1951 and it was found that there were 331 boys enrolled who had completed the blank in 1949. To determine whether or not the interests of those boys who moved from a predominantly rural environment to a metropolitan one had changed more than the interests of those boys who remained in a metropolitan environment, the 331 boys were divided into two groups. Those who graduated from high schools in Minneapolis, St. Paul, and their immediate suburbs were designated as a "metropolitan" group ( $N=250$ ). The second group, all of whom had graduated from high schools in cities of less than 20,000, were designated as a "non-metropolitan" group ( $N=81$ ).

A random sample of 125 boys was chosen from the metropolitan group. These boys plus the 81 non-metropolitan group were contacted in the spring of 1951 and asked to complete the Strong blank; 182, 88 per cent, complied. One blank

was unusable so that the scores of 181 boys were used in the study.

The minimum time between test and retest was two years and the maximum time did not exceed 2.5 years. The mean ages, at the time of the retest, of the metropolitan and non-metropolitan groups were 19.7 and 19.9 respectively. This difference was not statistically significant ( $P>.05$ ).

The median high school percentile rank of the metropolitan boys was 74.1 and that of the non-metropolitan boys was 79.9. The mean ACE Psychological Examination score of the metropolitan group was 119.69 and the mean for the non-metropolitan group was 122.06. These differences were not statistically significant ( $P>.05$ ).

The tests and retests for the 181 subjects were scored on 44 occupational keys and for Interest Maturity, Occupational Level, and Masculinity-Femininity. Also, using Darley's criteria (1), judgments of patterns were made for the eleven occupational interest groups. All judgments were made independently by two persons and in those cases where disagreement was found a third person made a third independent judgment. When more than two judges were needed the pattern was designated as that on which two of the three judges were in agreement. Thus, each of the eleven interest groups for each subject was scored as being a primary, secondary, tertiary, or "no" pattern. The third judge was needed for approximately five per cent of the judgments made.

### Results

*Permanence of Mean Scores.* One way of measuring the permanence of scores is to determine the stability of means between administrations. This was done separately for the 111 metropolitan and 70 non-metropolitan boys. The means and variances of the standard scores for 44 occupational and 3 non-occupational scales are given in Table 1.<sup>3</sup>

The significance of the difference between the test and retest means for each key was

<sup>3</sup> Table 1 has been deposited with the American Documentation Institute. Order Document No. 4239 from the ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress, Washington 25, D. C., remitting in advance \$1.25 for 35 mm. microfilm or \$1.25 for 6×8 in. photocopies. Make checks payable to Chief, Photoduplication Service, Library of Congress.

<sup>1</sup> This paper is based upon a portion of a Ph.D. thesis submitted to the graduate faculty of the University of Minnesota. The author wishes to acknowledge the guidance of his advisor, Dr. Willis E. Dugan.

<sup>2</sup> Now at Arkansas Polytechnic College.

tested by means of the *t* test, taking the correlation into account. Since the *F* test showed that the variances were significantly different in some cases, the assumption of homogeneity of variances did not hold in all cases; this is indicated in Table 1.

There was a significant difference between the means of the two administrations at the .01 level on 24 scales for the non-metropolitan group and on 26 scales for the metropolitan group. Twenty-two of these scales were common to the two groups. The significant changes in means between the test and retest, as shown in Table 1, were in both the positive and the negative direction. The direction, in all instances where there was a significant difference between administrations at the .01 level, was the same for the metropolitan and non-metropolitan groups.

The interest group which showed the largest and most consistent changes was Group V. All the scales within this group showed a significant increase in mean score for both the metropolitan and non-metropolitan boys. Of the non-occupational scales, Interest Maturity was the only one which changed significantly. As would be expected, the mean on this scale increased for both the metropolitan and non-metropolitan boys.

To determine whether or not there was a difference between the metropolitan and non-metropolitan groups with respect to stability of mean scores the means of the difference scores (test minus retest) were compared for each scale. When the variances were found to be homogeneous by means of the *F* test, the *t* test was used. When the variances were not homogeneous, the approximate method proposed by Cochran and Cox to test the hypothesis of equality of means with no hypothesis about the population variance was used (2). None of the differences between the means of the difference scores of the two groups were significant at the .01 level; three (Aviator, Vocational Agriculture Teacher, and Sales Manager) were found to be significantly different at the .05 level.

*Test-Retest Correlation.* The test-retest scores for each of the scales were plotted and in all cases the relationship between test and retest appeared, by inspection, to be linear.

Table 2  
Correlations Between Two Administrations of the Vocational Interest Blank for 111 Metropolitan and 70 Non-Metropolitan Males

Interest Group	Scale	Metropolitan Males	Non-Metropolitan Males
I	Artist	.72	.71
	Psychologist	.77	.67
	Architect	.75	.68
	Physician	.75	.64
	Osteopath	.71	.63
	Dentist	.71	.60
	Veterinarian	.74	.61
II	Mathematician	.59	.66
	Physicist	.78	.69
	Engineer	.78	.79
	Chemist	.79	.71
III	Production Manager	.62	.76
IV	Farmer	.85	.76
	Aviator	.78	.81
	Carpenter	.80	.71
	Printer	.61	.60
	Math. Phys. Science Tchr.	.72	.61
	Industrial Arts Tchr.	.79	.74
	Voc. Agri. Tchr.	.81	.68
	Policeman	.69	.60
	Forest Service Man	.80	.66*
	YMCA Physical Director	.72	.66
V	Personnel Director	.60	.49
	Public Administrator	.62	.45
	YMCA Secretary	.62	.60
	Soc. Science H. S. Tchr.	.68	.69
	City School Supt.	.70	.69
	Minister	.67	.70
	Musician	.68	.76
	CPA	.63	.60
	Senior CPA	.65	.62
	Accountant	.72	.66
VIII	Office Man	.66	.67
	Purchasing Agent	.72	.77
	Banker	.75	.64
	Mortician	.74	.74
	Pharmacist	.54	.59
	Sales Manager	.77	.62
	Real Estate Salesman	.72	.60
	Life Insurance Salesman	.79	.69
	Advertising Man	.75	.68
	Lawyer	.73	.73
IX	Author-Journalist	.70	.74
	President Mfg. Concern	.67	.60
	Interest Maturity	.66	.61
	Occupational Level	.70	.57
	Masculinity-Feminity	.76	.87

\* Difference significant at the .05 level.

Product moment correlation coefficients were then computed between the test and retest for each of the scales. These were computed separately for the metropolitan and non-metropolitan group. The correlations are given in Table 2.

None of the observed differences between the test-retest correlations for the metropolitan and non-metropolitan groups were found to be significant at the .01 level. Only one, that for Forest Service Man, was found to be significant at the .05 level.

*Permanence of Letter Grade Scores.* To get a measure of permanence in terms of letter grade scores, the change in letter grades between test and retest was determined. For each Strong scale a tabulation was made of the letter grade obtained on the retest for each letter grade received on the test. Since the comparisons of mean scores and of correlations for the metropolitan and non-metropolitan groups indicated that the two groups were similar with respect to permanence of scores, the metropolitan-nonmetropolitan classification was not retained for this part of the study. The two groups were pooled and treated as a single sample.

Table 3 gives the amount of change in letter grades between test and retest when all occupational scales are summed. Because of space limitations a breakdown by individual keys is not included here. For such a detailed breakdown see Stordahl (4).

A chi-square test for independence of letter grade and permanence was made by summing the letter grades over all scales and classifying the change in letter grades between test and retest into two categories—"identical" and "not identical." The hypothesis of independence of permanence and letter grade was rejected ( $P < .001$ ). Table 3 indicates that on the average, C grades were the most stable, 68 per cent of the C grades on the first test being C grades on the second test. The second most stable letter grade was A, with 60 per cent of the letter grades being identical on the test and retest. The intermediate letter grades were less stable. By combining the letter grades so that B included B +, B, and B -, and C included C + and C, it was found that 73 per cent of the C grades on the

Table 3

Change in Letter Grade Scores on the Vocational Interest Blank for 181 Boys Tested as High School Seniors and Retested Two Years Later as College Students

Test		Retest						
Letter Grade	N	% C	% C+	% B-	% B	% B+	% A	Total
A	804	2	3	6	10	19	60	100
B+	761	4	6	13	21	26	30	100
B	1,106	8	12	20	23	20	17	100
B-	1,394	19	16	24	20	12	9	100
C+	1,300	31	23	22	15	6	3	100
C	2,599	68	15	9	5	2	1	100

test remained C grades on the retest and that 59 per cent of the A and 59 per cent of the B grades remained constant.

*Permanence of Interest Patterns.* The permanence of interest patterns over the two year period is summarized in Table 4. Here, as for the letter grades, the data are presented for the metropolitan and non-metropolitan groups combined and for all interest groups combined.

Chi-square was used to test the independence of interest pattern and permanence by summing the patterns over all groups and classifying the change between tests as "identical" and "not identical." The hypothesis of independence was rejected ( $P < .001$ ).

As can be seen from Table 4, the primary and "no pattern" patterns were found to be the most stable with 58 per cent of the pri-

Table 4

Change in Interest Patterns on the Vocational Interest Blank for 181 Boys Tested as High School Seniors and Retested Two Years Later as College Students

Test		Retest			
Pattern	N	% N	% T	% S	% P
P	229	13	12	17	58
S	210	28	20	28	24
T	274	45	19	16	20
N	1,278	81	9	6	4



mary patterns on the first test being primary patterns on the retest and 81 per cent of the "no patterns" being identical on the retest. The secondary and tertiary patterns were less stable.

*Permanence of Individual Profiles.* The stability of individual profiles was determined for each of the 181 boys. Kendall's (3) coefficient of concordance,  $W$ , was used as a measure of stability. The coefficient,  $W$ , is based on the method of ranks. It is related to Spearman's  $\rho$ . The 44 occupational scales were used in computing this coefficient; the non-occupational scales were not included. When there were ties in rank, the median rank was assigned to each of the tied scales.

The median coefficient of concordance for the metropolitan group was .87 and the median for the non-metropolitan group was .86. Since  $W$  has a direct relationship to Spearman's  $\rho$ , these figures can also be expressed in terms of  $\rho$ ; the median  $\rho$ s being .74 and .72. All but nine of the coefficients for the metropolitan group and six for the non-metropolitan group were found to be significantly greater than zero.

The homogeneity of the frequency distributions of coefficients of concordance for the metropolitan and non-metropolitan groups was tested by the Brandt-Snedecor chi-square method (2). The two distributions were found to be homogeneous ( $P > .05$ ).

### Discussion

In this study a substantial relationship was found to exist between the interest scores received as high school seniors and as college sophomores. This relationship was, however, far from being a perfect one and large individual differences in stability were evident.

The scores of the metropolitan and non-metropolitan boys were quite homogeneous with respect to permanence of interest scores. The writer hypothesized that if the boys' interests had not as yet become stabilized that some difference between these groups might be found. Assuming that interests are largely determined by one's experiences, it was thought that the change in environment for

the non-metropolitan boys might cause a greater change in their scores than would be found for the metropolitan boys, whose environment remained relatively constant. Such a difference was not found.

No attempt has been made to make a direct comparison between the results of this study and the previous research of Strong and others. Such a comparison would be difficult since most previous research has been based on the original form of the blank whereas the revised form was used in the present study. However, since Strong (5, 6, 7) has reported some data on permanence with the revised keys and since the original keys were very similar to the revised, some general comparisons can be made.

The results of the present study, with respect to permanence as measured by mean scores, correlation, and permanence of individual profiles, are not greatly divergent from previous investigations. The results also support Strong's conclusion that we can place the greatest confidence in C letter grade ratings. Theoretically, as Strong has indicated, the A and C ratings should be the most stable as they cover a wider range of scores than the B rating. The fact that the A ratings were found to be no more stable than the B can probably be accounted for by the relatively small number of A ratings and the tendency for them to be low A ratings.

Although no previous studies have considered permanence in terms of interest patterns, counselors may find this the best way to look at permanence of scores. The evidence indicates that the counselor can place the most confidence in primary patterns and "no patterns" since these are apparently more stable than secondary and tertiary patterns. This is especially true when no pattern exists in an interest group.

### Summary

A sample of 181 males, 111 from a large metropolitan area and 70 from non-metropolitan areas, who had completed Strong's Vocational Interest Blank as high school seniors were retested two years later as college students. The tests and retests for the

181 boys were scored on 44 occupational keys and for Interest Maturity, Occupational Level, and Masculinity-Femininity. The test-retest scores on the 47 scales were compared in several ways to secure an estimate of the stability of scores over the two-year period. The following measures of test-retest stability were used: permanence of mean scores, test-retest correlation, permanence of letter grade scores, permanence of interest patterns, and stability of individual profiles.

A substantial relationship was found to exist between the interest scores received as high school seniors and as college sophomores. The metropolitan and non-metropolitan boys were quite homogeneous with respect to permanence of Strong Scores.

*Received December 30, 1953.*

## References

1. Darley, J. G. *Clinical aspects and interpretations of the Strong vocational interest blank*. New York: Psychological Corp., 1941.
2. Johnson, P. O. *Statistical methods in research*. New York: Prentice-Hall, 1949.
3. Kendall, M. G. *The advanced theory of statistics*, Vol. I. (4th ed.) London: Charles Griffin & Co., 1948.
4. Stordahl, K. E. *The stability of Strong Vocational Interest Blank patterns for pre-college males*. Unpublished doctor's dissertation, University of Minnesota Library, 1953.
5. Strong, E. K., Jr. Nineteen year follow-up of engineer interests. *J. appl. Psychol.*, 1952, 36, 65-74.
6. Strong, E. K., Jr. Permanence of interest scores over 22 years. *J. appl. Psychol.*, 1951, 35, 89-91.
7. Strong, E. K., Jr. *Vocational interests of men and women*. Stanford: Stanford Univer. Press, 1943.

# Adolescent Vocational Interests and Later Occupation

Phyllis Rosenberg Levine

*Jewish Vocational Service, Cleveland, O.*

and

Richard Wallen<sup>1</sup>

*Western Reserve University*

Although there is evidence that scores on the Kuder Preference Record differentiate among occupational groups (e.g., 2, 3, 5), few investigators have attempted to determine the relationship between such scores and the occupation entered at a later date. Barnette (1) presents data indicating that Kuder scores are related to occupational satisfaction several years after counseling, but his subjects were young adults at the time of advisement. Since many counselors deal with adolescent youth, it is of some interest to find out whether Kuder scores obtained during the late adolescent period are related to actual occupations entered subsequently.

The present paper reports a follow-up study of boys counseled at the Cleveland Jewish Vocational Service during the years 1943, 1944, and 1945. This study attempted to determine whether a relationship existed between Kuder Preference Record (Form B) scores at the time of counseling and the occupation engaged in at the time of the study (1952).

## Subjects

Questionnaires designed to elicit information about current occupation were sent to 215 men who had been counseled and tested at JVS during the years 1943-1945. The mailing list included all eleventh and twelfth-grade males tested during that period and all tenth-grade males tested in 1943 and 1944. The original letter, a reminder postcard, and a follow-up letter yielded questionnaire returns from 58 per cent of the mailing list. Of the total group, six per cent were known to be in military service, seven per cent had moved to unknown addresses, and one per cent was deceased. No information of any kind was returned for 28 per cent of the group.

<sup>1</sup> This paper is based on a portion of a thesis submitted in partial fulfillment of the requirements for the M.A. degree at Western Reserve University by the first-named author and supervised by the second author.

In order to determine the existence of bias in the sample that returned questionnaires, several comparisons were made between the respondent and non-respondent groups. The mean age of the 124 respondents at time of original testing was 16 years and 7 months and that of the 64 non-respondents was 16 years 10 months, 3 months higher. The difference is statistically insignificant ( $t = .39$ ). Scores on the American Council Examination (High School Form, 1942) were available for 44 of the respondents and for 23 of the non-respondents. The means of the two groups were not significantly different for L, Q, or Total scores. (For Total scores,  $t = .88$ .) When mean scores on the separate scales of the Kuder were compared, only one difference was significant at the five per cent level. That difference occurred between mean scores on the musical scale, and the respondents were significantly lower in measured musical interest. These comparisons suggest that the respondent group is a fairly unbiased sample of the total group to whom questionnaires were mailed.

Further data on the respondent group were obtained from the questionnaires. In terms of educational achievement, it is clear that our respondents are not representative of the general male population and are probably not representative of males seen at the agency. Of the respondents, 79 per cent indicated that they had completed at least four years of college. Less than two per cent had not finished high school. This substantial educational attainment is probably due in part to the financial assistance given veterans during the period covered by the study, but it may also reflect the family educational aspirations held by the clients of this agency.

## Procedure

The questionnaire consisted of three items asking for identifying data, one item on amount of education, two items requesting data about present occupation and length of time it had been engaged in, four items dealing with job satisfaction, and one item about the estimated influence of JVS counseling on occupational choice. The data on job satisfaction are not analyzed in the present report, since few respondents indicated dissatisfaction with their current occupation.



The first step in the treatment of the data was to classify the reported occupations of the respondents. Using the Kuder manual (3) as a guide, the occupations were coded as belonging to one or more of the nine Kuder interest areas. For example, mechanical engineering was classified as belonging to the mechanical, computational, and scientific interest groups. In most cases the reported occupation was classified according to its listing in the Kuder manual. A subjective judgment had to be made in a small number of cases: T-V producer was classed as persuasive, artistic, literary, and musical; graduate student in international relations was classed as persuasive, literary, and social service; business men, executives, and those who were self-employed were included in the persuasive interest group. The occupational interest classification was entered on three by five cards containing other data about the subjects, so that tabulation could be done directly from the cards.

Seven cases were eliminated from the respondent group at this point. They were either unemployed or were in undergraduate college. The final group of respondents used in this study, then, numbered 117.

For each Kuder scale, the total group of respondents was divided into two sub-groups: those in occupations belonging to that interest area and all others. Mean Kuder raw scores and standard deviations were computed for each of these sub-groups, and a *t*-test was applied to the differences between the means.

### Results

Table 1 summarizes the results of the statistical analysis. Taking the mechanical interest scale as an example, Table 1 reads as follows: Of the entire respondent group, 26 were currently occupied in jobs involving mechanical interest, and 91 were in jobs that did not require this interest. The mean mechanical interest score earned by the mechanically occupied group, seven to nine years earlier, was 81.6. The mean score of those now in other kinds of jobs was 69.5. The difference between the mean scores of the two groups is significant at the five per cent level of confidence as shown by the *t* value of 2.46.

Only mean scores are presented for the artistic and musical scales, since few subjects reported occupations involving these interests. On both of these scales, however, the differences are in a direction consistent with those found for the other scales.

For six of the remaining seven scales, the data show that men currently engaged in oc-

Table 1  
Comparisons of Kuder Preference Record Scores Made  
Seven to Nine Years Ago by Men Engaged in  
Occupations Related to an Interest Area  
and by Those Engaged in Other  
Occupations

Kuder Scale	N	M	S.D.	<i>t</i>
<i>Mechanical</i>				
Occupied	26	81.6	21.5	2.46*
Others	91	69.5	22.8	
<i>Computational</i>				
Occupied	34	43.6	11.5	3.55**
Others	83	35.5	9.9	
<i>Scientific</i>				
Occupied	30	86.1	11.1	5.39**
Others	87	71.1	17.5	
<i>Persuasive</i>				
Occupied	53	78.8	17.2	4.48**
Others	64	64.5	17.0	
<i>Artistic</i>				
Occupied	5	47.6	—	
Others	112	40.2	—	
<i>Literary</i>				
Occupied	13	65.5	13.6	3.02**
Others	104	52.7	14.9	
<i>Musical</i>				
Occupied	2	31.0	—	
Others	115	19.0	—	
<i>Social Service</i>				
Occupied	10	75.0	18.3	1.59
Others	107	64.9	17.5	
<i>Clerical</i>				
Occupied	24	59.0	13.2	3.93**
Others	93	49.7	15.4	

\* Significant at the 5 per cent level of confidence.

\*\* Significant at the 1 per cent level of confidence.

cupations involving those interests made significantly higher mean scores than did men in other occupations.

The failure of the men engaged in social service occupations to show a significantly higher social service score than those in other occupations probably reflects an inadequacy of our sample rather than an inadequacy of the scale. Eight of the ten cases classed as engaged in a social service occupation were students in professional school. Only one of these was in a graduate school of social work.

Three others were engaged in the graduate study of liberal arts subjects, three were in medical school, and one was in dental school. The composition of this sub-group, then, does not provide a satisfactory sample of persons in this occupational interest area.

These results provide evidence that interest scores on the Kuder Preference Record are positively related to occupations entered seven to nine years later. Further, they indicate that interests have been sufficiently organized by the time the last few years of high school are reached to provide one basis for estimating future occupational activity.

### Discussion

Several considerations should be kept in mind in interpreting the results of this study. In the first place the interest test was administered as part of a total counseling service. Aptitude and achievement tests were used along with interest tests to provide a basis for personal interviews. It could be argued that the decisions arrived at during the entire counseling process largely determined the occupation entered seven to nine years later. If the Kuder scores influenced counseling decisions, then the relationship found in this study could be due, not to the persistence of adolescent interests, but to the persisting effects of counseling based on adolescent interests. While our data cannot settle this issue definitively, several facts argue against the belief that counseling decisions alone can account for the relationship between measured interests and occupational entry. For one thing, Strong (4) has presented findings that show persistence of interests over a long period of time. His original test data were apparently not collected in a counseling situation, so that the persistence of interests he found could not be attributed to the counseling process. For another thing, our respondents themselves did not attribute a great deal of influence to the decisions arrived at in the counseling process. When asked whether the suggestions made by the JVS influenced their occupational plans, only 35 per cent said "yes," 44 per cent said "no" and 20 per cent could not recall any influ-

ence. Although the recall of counseling influence is not the sole valid measure of the existence of influence, our data certainly do not support the view that it determines occupational entry to a greater extent than the persistence of interests.

A further consideration in interpreting our findings concerns the effect of military service and government assistance to veterans in school. Perhaps military service had the effect of disrupting the normal peacetime paths to occupational entry. Our results would then show the minimal relationship between adolescent interest and later occupation. On the other hand, our respondents may have been enabled to enter preferred occupations to a greater extent than is usually true, because the veterans' benefits helped them to continue their education. The most that can be said on this point is that our findings need to be supported by data collected during a period free from the special influences created by wartime mobilization and a postwar economy.

### Summary and Conclusions

In order to discover whether a significant relationship existed between adolescent interests and later occupational choice, a questionnaire was mailed to 215 men who had been counseled seven to nine years earlier during the latter portion of their high school careers. Usable information on current occupation was obtained from 117 of those on the mailing list. Comparisons of the respondents with the non-respondents indicated no difference with respect to age at time of counseling, intelligence, and mean scores on eight of the nine Kuder Preference Record scales. Reported occupations were classified in accordance with the interests they involved as presented in the Kuder manual.

For six of the Kuder interest areas, men currently engaged in a related occupation made significantly higher scores seven to nine years ago than did men engaged in unrelated occupations. The three remaining interest areas (artistic, musical, and social service) did not yield clear-cut results because of the inadequacies of the sample.

We conclude that interests measured by the Kuder Preference Record in adolescence are positively related to occupation engaged in seven to nine years later.

Received November 13, 1953.

#### References

1. Barnette, W. L., Jr. *Occupational aptitude patterns of selected groups of counseled veterans*. *Psychol. Monogr.*, 1951, 65, No. 5 (Whole No. 322).
2. Hahn, M. E. and Williams, Cornelia T. The measured interests of Marine Corps Women Reservists. *J. appl. Psychol.*, 1945, 29, 198-211.
3. Kuder, G. F. *Revised manual for the Kuder Preference Record, Vocational, Form B*. Chicago: Science Research Associates, 1946.
4. Strong, E. K., Jr. Permanence of interest scores over 22 years. *J. appl. Psychol.*, 1951, 35, 89-91.
5. Triggs, Frances O. The measured interests of nurses: A second report. *J. educ. Res.*, 1948, 42, 113-121.



## The Degree to Which Colors (Hues) Are Associated with Mood-Tones<sup>1</sup>

Lois B. Wexner

*Division of Education and Applied Psychology, Purdue University*

The literature is replete with statements concerning the relation of color and emotional states or feeling-tones, but there is a dearth of experimental investigation to support these statements. In a recent study by Odber, Karwoski, and Eckerson (8), regarding the associations of color and mood, it was found that some colors were more often chosen to go with certain groups of words describing mood, such as red with exciting, orange with gay, yellow with playful, green with leisurely, blue with tender, purple with solemn, and black with sad. Two shortcomings of this study, however, are first, that the groups of words (represented above by exciting, gay, etc.) included words which could in no way be considered to mean the same thing, such as the "playful" list, which included humorous, whimsical, fanciful, quaint, sprightly, delicate, light, and graceful. Thus one subject may be reacting to one particular word in the list, and another, to an entirely different one. And, second, a partially "forced" method was used to fit the moods to a color-circle (arranged according to wave-length). For instance, gay is reported to "go with" orange, but in reality orange was chosen only 16 times, whereas red was chosen 62 and yellow 27 times. Further, the authors' judgment appears to be the only method used to choose which colors went with which moods. Thus, although the numerical results are published, a clear-cut statistical interpretation is lacking. Other studies (1, 2, 3, 9, 10, 13) report the association of color and moods, as determined by various methods including objective impressions, clinical observation, and introspection.

### Purpose

The purpose of this investigation is to determine to what degree colors (hues) are associated with mood-tones. The hypothesis to be tested is that there is a positive relation between certain colors and mood-tones.

<sup>1</sup> Grateful acknowledgment is made to James A. Norton, Jr., for his helpful suggestions in the use of statistical techniques, and to Joyce Block, Malcolm Robertson, and Henry Wexner, who served as judges.

### Procedure

The mood-tones used in this experiment were arbitrarily selected as a fairly representative group. Originally, twelve words were chosen, i.e., exciting, secure, distressed, tender, protective, despondent, calm, dignified, cheerful, defiant, powerful, and sensuous. Then a list of 164 adjectives was prepared, including moods reported in the literature, synonyms of those words as well as those listed above, and other words the writer believed might be useful. The original twelve words were presented to four judges, with the list of adjectives. The judges (two of whom were male and two female) were requested to choose words from the list of adjectives which they felt meant the same as the "mood-tone" words. They were allowed to use words more than once if they wished. Then, the mood-tone words were listed together with their synonyms as unanimously chosen by the four judges. Since the judges did not agree on the meaning of sensuous, this word was not included in the experiment. The final groups of mood-tones are as follows: exciting, stimulating; secure, comfortable; distressed, disturbed, upset; tender, soothing; protective, defending; despondent, dejected, unhappy, melancholy; calm, peaceful, serene; dignified, stately; cheerful, jovial, joyful; defiant, contrary, hostile; and powerful, strong, masterful.

The subjects consisted of 94 students in a course of beginning General Psychology, of which 48 were female and 46 were male. The subjects, in three groups, were presented with an instruction sheet containing the word groups as above and the following directions:

The following groups of words are meant to represent feelings, or mood-tones. It is thought that certain colors tend to "go with" various mood-tones, and this is an attempt to determine to what extent this may be true. Please select the one color, of the colors on the charts, that you feel best represents the feelings described by the following word groups. All the colors need not be used, and colors may be used more than once. Be sure to select a color for each group, even though it may seem difficult to find a color to fit the mood-tone. Usually your first impression would be the best one, if in doubt.

Eight colors, yellow, orange, red, purple, brown, blue, black, and green, in the form of  $8\frac{1}{2} \times 11$  inch pieces of art paper mounted on  $30 \times 40$  inch pieces of light-gray cardboard, were randomly arranged at the front of the room. It should be noted that no mention of color names was made by the experimenter. This was in order to avoid associations to color stereotypes

and to assure that the colors as chosen were the particular shades presented to the subjects, in an attempt to insure uniformity of shade. It might further be noted that there was no difficulty in determining which colors the subjects intended to indicate.

Chi-square tests for any possible sex differences in color association to mood-tones were made. No significant differences were found to exist.

Since there were no significant sex differences, the frequencies from the two sexes were combined into one set for further study. Then, for each mood-tone, a chi-square test was made to test whether or not the colors differed significantly in frequency of association with that mood-tone. These chi-squares were significant in all cases. (A five per cent significance level was used.) Thus it is demonstrated that some colors are more often associated with a given mood-tone than others.

The next step was to determine which particular colors were most often associated with a given mood-tone. For this purpose, Tukey's (14) procedure for accomplishing multiple comparisons among a set of observed means was adapted to make multiple comparisons among a set of observed frequencies in mutually exclusive categories (7). The essential nature of this adaptation was to use the inverse sine transformation upon the observed proportions. The error variances of such transformed proportions are given by the theory of the transformation (4). In all cases a significance level of five per cent was used.

### Results

The following results were obtained. For each mood-tone, the colors are grouped (A, B, C, etc.) according to the results of the multiple comparisons tests. The interpretation of these groups is as follows: colors in the same group are associated with the mood-tone significantly more often than colors in groups below them, and significantly less often than colors in groups above them. Colors in the same group do not differ significantly from each other in frequency of association with the mood-tone.

#### *Exciting, stimulating.*

Group	Color	Frequency
A	Red	61
B	Yellow	12
	Orange	11
C	Green	4
	Purple	4
	Black	2
	Blue	2
	Brown	0

#### *Secure, comfortable.*

Group	Color	Frequency
A	Blue	41
B	Brown	23
	Green	18
C	Yellow	8
D	Orange	2
	Black	2
	Red	0
	Purple	0

#### *Distressed, disturbed, upset.*

Group	Color	Frequency
A	Orange	34
B	Black	16
C	Purple	10
	Brown	9
	Green	8
	Red	7
	Yellow	5
	Blue	5

#### *Tender, soothing.*

Group	Color	Frequency
A	Blue	41
B	Green	24
C	Yellow	11
	Purple	9
	Brown	6
D	Orange	2
	Black	1
	Red	0

#### *Protective, defending.*

Group	Color	Frequency
A	Red	21
	Brown	17
	Blue	15
	Black	15
	Purple	14
B	Green	5
	Orange	4
	Yellow	3

#### *Despondent, dejected, unhappy, melancholy.*

Group	Color	Frequency
A	Black	25
	Brown	25
B	Purple	11
	Blue	11
	Green	9
	Yellow	5
	Orange	4
C	Red	0

*Calm, peaceful, serene.*

Group	Color	Frequency
A	Blue	38
	Green	31
B	Yellow	8
	Purple	7
	Orange	4
	Brown	3
	Black	3
C	Red	0

*Dignified, stately.*

Group	Color	Frequency
A	Purple	45
B	Black	30
C	Blue	9
	Brown	6
D	Red	3
	Orange	1
	Yellow	0
	Green	0

*Cheerful, jovial, joyful.*

Group	Color	Frequency
A	Yellow	40
B	Red	20
C	Orange	14
	Green	11
	Blue	7
D	Purple	2
	Brown	0
	Black	0

*Defiant, contrary, hostile.*

Group	Color	Frequency
A	Red	23
	Orange	21
	Black	18
B	Brown	11
	Purple	9
	Yellow	5
	Green	5
C	Blue	2

*Powerful, strong, masterful.*

Group	Color	Frequency
A	Black	48
B	Red	23
C	Purple	8
	Blue	6
	Brown	4
	Orange	3
	Yellow	1
	Green	1

## Discussion

In general, the results of this investigation tend to support the color-mood studies as reported in the literature. It should be noted, however, that the association of some mood-tones with certain colors is more clear-cut than others. For instance, in some cases one color "goes with" a mood-tone significantly more often than does any other color (of the particular shades of colors used in this experiment). Red is more often associated with exciting-stimulating, blue with secure-comfortable, orange with distressed-disturbed-upset, blue with tender-soothing, purple with dignified-stately, yellow with cheerful-jovial-joyful, and black with powerful-strong-masterful. On the other hand, there is no statistically significant difference between certain colors in their association with certain other mood-tones, such as red, brown, blue, black, and purple with protective-defending; black and brown with despondent-dejected-unhappy-melancholy; blue and green with calm-peaceful-serene; and red, orange, and black with defiant-contrary-hostile.

Since there appears to be fairly consistent agreement among the studies on this subject, it is appropriate to suggest possible contributing factors, although it is not the purpose of this paper to investigate this particular aspect of the problem. In addition to the cultural factor which no doubt plays an important part in the associations of colors with certain mood-tones, there seems to be the possibility of the existence of biological determinants. Guilford (6) states that experimental results "point very strongly to a basic communality of color preferences among individuals. This communality probably rests upon biological factors, since it is hard to see how cultural factors could produce by conditioning the continuity and system that undoubtedly exists." Goldstein (5) is more explicit in setting forth physiological effects of color on the human organism, and indicates that patients, exposed to various colors such as large sheets of colored paper, change the position of the arms in different directions, according to the color to which they are exposed; that color influences the speed of volitional movements; and that seen and felt distances and time intervals and



weights are judged differently under the influence of different colors. He finds, furthermore, that green favors performance in general, in contrast to red, and feels that these different effects correspond to very definite, but different, total behavioral attitudes, which find their expression very clearly in the subject's reports of the mood corresponding to the various colors.

In addition to the part played by learning in the cultural and biological determinants of associations of colors with certain mood-tones, there may be an additional factor which should be included, in the form of particular learning situations, which may affect individuals and/or groups. An experiment in support of this type of contribution was done by Staples and Walton (12).

The foregoing are merely suggested as possible contributing factors to color and mood association, and the need for additional experimental work in this area is obvious.

With regard to the present investigation, it would seem possible, and even likely, that in a similar experiment different results might be obtained if different shades of the same colors were used. For instance, in a discussion with a group of the subjects after the data had been collected, the writer mentioned that she had expected purple to "go with" powerful. One of the subjects replied that the particular shade of purple was not deep and dark enough to be a "powerful" purple. Thus it would appear that extreme caution should be used in generalizing these findings to other shades of the same colors. However, because of the positive findings of this experiment, it would appear that useful information could be obtained by extending this type of investigation to other groups. Such information might possibly be of extensive value to both industrial and clinical psychologists.

### Summary

In an attempt to determine to what degree colors (hues) are associated with mood-tones, 94 subjects were presented with eight stimulus colors (red, orange, yellow, green, blue, purple, brown, and black) and a list of eleven moods (exciting-stimulating; secure-comfortable; distressed-disturbed-upset; tender-soothing; protective-defending; despondent-de-

jected-unhappy-melancholy; calm-peaceful-serene; dignified-stately; cheerful-jovial-joyful; defiant-contrary-hostile; and powerful-strong-masterful), the word selections of which had been unanimously agreed upon by four judges. No significant differences were found in color-mood association between male and female. It was found, however, that for each mood-tone certain colors were chosen to "go with" that mood-tone significantly more often than the remaining colors, and the results were stated.

Inasmuch as there was general agreement among studies concerning mood and color association, several possibilities for this were given, such being the influence of cultural, biological, and learning factors.

Received December 3, 1953.

### References

1. Birren, F. *Color psychology and color therapy*. New York: McGraw-Hill, 1950.
2. Buck, J. N. *Proceedings of the H-T-P workshop*. Richmond, Va.: Veterans Administration Hospital, 1950.
3. Chandler, A. R. *Beauty and human nature*. New York: Appleton-Century, 1934.
4. Eisenhart, C., Hastay, M. W., and Wallis, W. A. *Techniques of statistical analysis*. New York: McGraw-Hill, 1947.
5. Goldstein, K. *The organism*. New York: American Book Co., 1939.
6. Guilford, J. P. There is system in color preferences. *J. opt. soc. Amer.*, 1940, 30, 455-459.
7. Nair, K. R. The distribution of the extreme deviate from the sample mean and its studentized form. *Biometrika*, 1948, 35, 118-144.
8. Odber, H. S., Karwoski, T. F., and Eckerson, A. B. Studies in synesthetic thinking: I. Musical and verbal associations of color and mood. *J. gen. psychol.*, 1942, 26, 153-173.
9. Risler, J. L'influence psychologique de la lumiere. (The psychological influence of light.) *Cour. med.*, 1927, 77, 40-42.
10. Rogers, Marian E. *A study of color preferences*. Unpublished master's thesis, Purdue University, 1950.
11. Snedecor, G. W. *Statistical methods* (4th ed.). Ames: Iowa State College Press, 1946.
12. Staples, R. and Walton, W. E. A study of pleasurable experiences as a factor in color preference. *J. genet. psychol.*, 1933, 43, 217-223.
13. Tatibana, Y. Color feelings of the Japanese. I. The inherent emotional effects of colors. *Tohoku psychol. fol.*, 1937, 5, 21-46.
14. Tukey, J. W. Comparing individual means in the analysis of variance. *Biometrics*, 1949, 5, 99-114.

## Readability of Mathematical Tables\*

Miles A. Tinker

*University of Minnesota*

Casual examination of several mathematical or statistical tables will reveal great variation in the typographical arrangements employed. From table to table the reader may find variation in type size, type face, use of additional leading at periodic intervals, number of decimal places employed, etc. To a reader with some background in scientific typography, it is obvious that some of these factors should influence the readability of the tables. Since a particular mathematical table may be put to a number of different uses by workers or students in a variety of scientific fields, it would seem that arbitrary choice of a specific typographical arrangement for use in a certain field is not the most important factor to consider, particularly with tables of squares, cubes, square roots, and cube roots which are widely used. In some tables, economy of space seems to be the sole consideration with no attention to readability factors. Where readability (or legibility) is mentioned, as by Milne (3), choice of typography depends upon opinion rather than upon experimental findings.

Actually, there has been no experimental work done on the readability of mathematical tables. A few related findings in specialized kinds of reading may be cited: Baird (2) studied the legibility of a telephone directory. He found that  $\frac{1}{2}$  point leading between lines was 13 per cent more efficient in terms of time taken to find a number than when set solid. He also found that indenting every other line in the directory increased only slightly (probably not significantly) the speed and accuracy of locating telephone numbers in comparison with an even alignment of names. Scott (5) had subjects read two pages of a railroad time-table, each set up in light-faced small type and heavy-faced large type. The large heavy-faced type was read faster and with considerably fewer errors.

\* The writer is grateful to the University of Minnesota Graduate School for research grant to finance this study.

Size of type rather than heaviness of type face may have been the important factor. After inspecting a number of mathematical tables, Babbage (1) expressed a preference for numerals of uniform height (modern) rather than those with ascenders and descenders (Old Style). In a report (4) of the Committee on Type Faces it is recommended, on the basis of collected opinions, that modernized Old Style numerals be used in mathematical tables. Reading the Old Style numerals is considered to produce less fatigue. Milne (3) also considers the Old Style numerical symbols, in which most of the characters have heads or tails, to be more legible than those of uniform height. Tinker (6) determined (a) the relative visibility of Modern and Old Style numerals by obtaining the average distance from the eyes at which the numerals could be read correctly; and (b) the speed and accuracy of reading the two kinds of numerals. The Old Style numerals, read in isolation, were slightly more visible (probability at the 2 per cent level), but in groups were much more visible (probability beyond the one per cent level). But Modern numerals in groups were read just as fast and just as accurately under normal reading conditions as the Old Style numerals. It was suggested that when numerals are printed in groups as in tables, that the Old Style numerals be used because they are perceived more easily (more visible).

The above citations merely suggest what might be more satisfactory in terms of size of type, leading, and type style. Since no experimenting with actual tabular materials has been done, the need for some exploratory investigation seems indicated. The purpose of the present study is to investigate the comparative readability of five mathematical tables in terms of the speed with which subjects can find the squares, square roots and cube roots of numbers. Tables were chosen which permitted comparisons between type



sizes, type faces, and arrangement of columns and rows of numerals.

### Materials and Procedure

The five published tables will be designated by the letters: A, B, C, D, and E. For purposes of comparison we will need a rather complete description of each table. Only tables which included squares, cubes, square roots and cube roots were used from each book (except Table A which did not include cubes and cube roots). Table 1 shows the columnar arrangements of the five mathematical tables.

*Table A* has a  $6 \times 9$  inch page. The numerals are printed in an 8 point Modern (all numerals same height) type set solid with successive groups of five entries down the columns separated by 8 point leading. The first column (No.) is in bold face and the remaining numerals ordinary light-face. Decimals are carried to three places. In the square column, there is  $\frac{1}{4}$  pica space between each set of two numerals along a line. Columns are separated by a 1 pica space with no rule. The paper is a good quality mat white and thick enough so that shadows from print on the reverse side do not show through.

*Table B* has a  $5\frac{3}{8} \times 8\frac{1}{2}$  inch page. The numerals are printed in an 8 point Old Style type (ascenders and descenders) set solid with successive groups of five entries down the columns separated by 8 point leading. The first column (No.) is in bold face and the remaining numerals in ordinary face. Decimals are carried to seven places. In the square column, the numerals along a line are grouped in twos as in Table A, and by threes in the cube column. There is  $\frac{3}{4}$  to 1 pica space plus a rule between various columns. The paper is good quality mat white and thick enough so that no shadows show through.

*Table C* has a  $3\frac{1}{8} \times 6\frac{3}{4}$  inch page. The numerals are printed in a 6 point Modern type, set

solid with groups of 10 entries down the columns separated by 6 point leading. There is no bold face type in this table. Decimals are carried to four places. There are no groupings into twos or threes along lines in the squares and cubes. There is a 1 pica space with no rule between columns. The paper is a good quality mat white and thick enough so that no shadows show through.

*Table D* has a  $3\frac{3}{4} \times 6\frac{3}{4}$  inch page. The numerals are printed in a 6 point Modern type, set solid with no grouping of entries down the columns, but each fifth entry down a column is in bold face which is only a little darker than the rest of the printing. Decimals are carried to 7 places in the square and cube roots. There are no groupings along lines into twos or threes in the squares and cubes. There is a  $\frac{1}{2}$  pica space plus a rule between columns. The mat grayish white paper is so thin that shadows from print on the reverse side show through enough to hinder discrimination of the numerals. The printed page impresses the reader as being crowded and difficult to read.

*Table E* has a  $4\frac{1}{8} \times 6\frac{1}{2}$  inch page. The numerals are printed in a 6 point Modern type, set solid with groups of 10 entries down the columns separated by 6 point leading. Numerals in the No. column are in bold face. Square root decimals are carried to four places, cube root to five places. There are no groupings along lines into twos or threes in squares and cubes. There is a  $\frac{1}{2}$  to 1 pica space between various columns plus a rule. The mat grayish paper is so thin that disturbing shadows show through from the print on the reverse side but these shadows are not as prominent as in Table D.

The typographical arrangements of these five tables permit a number of interesting readability comparisons: type face, A vs. B; type size, A vs. C; leading between grouping of entries down col-

Table 1  
Arrangement of Columns in Five Mathematical Tables

Table	Columnar Arrangement							
A	No.	Square	Square Root	No.	Square	Square Root	No.	Square Root
B	No.	Square	Cube	Fourth Power	$\frac{1}{\sqrt{N}}$	Square Root	Cube Root	$\frac{1}{N}$
C	No.	Square	Cube	Square Root	Cube Root	Circum.	Area	No.
D	No.	Circum.	Area	Square	Cube	Square Root	Cube Root	Reciprocal
E	No.	$\frac{1000}{\text{No.}}$	Square	Cube	Square Root	Cube Root	Circum. of Circle	Area of Circle



umns, D vs. E; arrangement of columns, various; etc. All tables have 50 entries per column except D which has approximately 75.

The experiment was carried out in a laboratory room with uniform illumination of 20 foot-candles. The subjects were 120 university students.

There were two general procedures: 1. The book was opened to page one of the table and on presentation of the number, the subject found the entry, turning pages where necessary, and read off the response. 2. The book was opened to the page of the table containing the number involved. Upon presentation of the number, the subject found the entry and read off the response.

This was done separately for the finding of squares, square roots, and cube roots. There were, therefore, six parts to the experiment. Twenty different subjects observed on each part. Materials (the five tables) and subjects were systematically rotated within each part to equate practice effects.

Ten numbers were looked up in each table by each subject. Upon arrival at the laboratory the subject was allowed to look over the five tables to become acquainted with them. He was then told that he would be presented with a number and that he was to look up and read off aloud the square (square root, cube root) as rapidly as possible. Two practice trials were given on each table just before it was used. The number to be looked up was presented typed on a 3 × 5 inch index card. The ten numbers to be looked up included one from each hundred up to a thousand (as 86, 141, 216, 324, 434, 538, 663, 728, 836 and 982). A different series of numbers was used for each table. Times were recorded in seconds and tenths of a second from presentation of the number (uncovered on table before the subject) to the beginning of the spoken response. All errors were tabulated. No information about results was given to a subject until the experiment was completed.

### Results

The basic data of this study are given in Table 2. Comparison of the lower with the upper half of the table reveals that much more time is taken to find squares, square roots and cube roots of a number when the book is opened at the beginning of the mathematical table (subject finds page) than when the book is opened to the page containing the item (subject given page). When mean scores for one-half of the subjects were compared with those for the other half the consistency of trends from mathematical table to table was high in each part of the experiment.

Table 2

Mean Time in Seconds Taken to Locate Squares, Square Roots and Cube Roots in Five Mathematical Tables

(N = 20 college students in each comparison, 120 in all)

Table	Squares		Square Root		Cube Root	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
Subject Finds Page						
A	5.18	.61	5.04	.49	—	—
B	5.24	.72	5.77	.86	5.20	.87
C	5.52	.70	5.70	.99	5.90	.86
D	6.34	.91	6.43	1.04	6.06	1.02
E	5.49	.71	6.30	1.08	6.06	.94
Page Given Subject						
A	2.99	.40	2.90	.37	—	—
B	2.74	.52	2.90	.66	2.77	.57
C	2.74	.54	2.92	.62	2.96	.64
D	3.71	.61	3.50	.86	3.41	.69
E	3.02	.66	3.06	.69	2.90	.61

Note: Original computations were carried to four decimal places.

Correlation of the ranks obtained from these scores ranged from .80 to 1.00. Tabulation of the errors revealed a high percentage of accuracy. In only two per cent of the item responses were there errors. There was little difference in error count from table to table although they were somewhat fewer in mathematical Table A. We may, therefore, concentrate our attention on the speed scores.

In Table 3 are listed differences and critical ratios for reading squares, square roots and cube roots of numbers when the subjects always started at page one of the mathematical table. In finding squares of numbers, starting always at the beginning of the mathematical table, times were significantly faster in Tables A, B, C and E than in D. With less certainty, time for A was faster than for C. Note also that A was better than E, and that B was better than C and E although the differences were not significant. There was no important difference between A and B or C and E. (The whole pattern of differences will be coordinated below under discussion.)

The data on significance of differences for finding square roots, starting at the beginning

Table 3

Differences Between Means in Seconds with Critical Ratios for Finding Squares, Square Roots and Cube Roots in Five Mathematical Tables When Subject *Finds Page*

Tables Compared	Squares		Square Roots		Cube Roots	
	Diff.	C.R.	Diff.	C.R.	Diff.	C.R.
A vs. B	+ .06	0.29	+ .73	3.32**	—	—
A vs. C	+ .34	1.65	+ .66	2.68**	—	—
A vs. D	+1.16	4.74**	+1.39	5.41**	—	—
A vs. E	+ .31	1.48	+1.26	4.77**	—	—
B vs. C	+ .28	1.25	— .07	0.24	+ .70	2.13*
B vs. D	+1.10	4.24**	+ .66	2.19*	+ .86	2.40*
B vs. E	+ .25	1.10	+ .53	1.72	+ .86	2.50*
C vs. D	+ .82	3.19**	+ .73	2.27*	+ .16	0.45
C vs. E	— .03	0.13	+ .60	1.84	+ .16	0.47
D vs. E	— .85	3.29**	— .13	0.39	.00	0.00

\* Significant at the 5 per cent level.

\*\* Significant at the 1 per cent level.

of the mathematical tables, reveal that Table A was significantly superior to all other Tables (B, C, D, E). With a lesser degree of significance (5 per cent level), B was better than D, and C than D. Also, B was better than E and C than E although the significance of the differences did not reach the 5 per cent level.

In locating cube roots, Table B was superior to C, D, and E. No other differences were significant.

Turning now to the situations in which the

mathematical tables were opened by the experimenter to the page containing the item to be located, the data on significance of differences for finding squares, square roots and cube roots are given in Table 4. In finding squares, Tables A, B, C and E are much better than D. B and C are somewhat better than A although not significant at the 5 per cent level. There is no important difference between A and E, or B and C.

The data for significance of differences for square roots with proper page given to the

Table 4

Differences Between Means in Seconds with Critical Ratios for Finding Squares, Square Roots and Cube Roots in Five Mathematical Tables When Subject is *Given Page*

Tables Compared	Squares		Square Roots		Cube Roots	
	Diff.	C.R.	Diff.	C.R.	Diff.	C.R.
A vs. B	— .25	1.71	.00	0.00	—	—
A vs. C	— .25	1.66	+ .02	0.12	—	—
A vs. D	+ .72	4.43**	+ .60	2.87**	—	—
A vs. E	+ .03	0.17	+ .16	0.92	—	—
B vs. C	.00	0.00	+ .02	0.10	+ .19	0.83
B vs. D	+ .97	5.42**	+ .60	2.48*	+ .64	2.67**
B vs. E	+ .28	1.49	+ .16	0.75	+ .13	0.59
C vs. D	+ .97	5.32**	+ .58	2.46*	+ .45	1.78
C vs. E	+ .28	1.47	+ .14	0.68	— .06	0.25
D vs. E	— .69	3.45**	— .44	1.79	— .51	2.08*

\* Significant at the 5 per cent level.

\*\* Significant at the 1 per cent level.

subject reveal that Tables A, B and C were much better than D. There were no other significant differences.

Data on significance of differences in looking up cube roots, page given, indicate that Tables B and E were considerably better than D. Other differences are unimportant.

### Discussion

The results obtained when the subject always started from the beginning page of the mathematical tables will be considered first. This is the kind of situation encountered by the reader in his customary use of tables of this kind. In practically every instance, a significantly greater time was required to locate squares, square roots and cube roots in Table D than in the other tables. In Table D, the type was smaller than in A or B; in D there was no additional leading to separate groups of items down columns to aid in following across rows in contrast to A, B, D and E; also, the paper was thinner in D than in A, B and C. It is possible that the arrangement of columns in D retarded finding the proper entry since it was necessary to skip over the first two columns next to the No. column. However, this may not be important since an analogous situation occurs for some entries in B and E. The fact that E is better than D for finding squares is probably due to the additional leading which separates groups of items down columns.

Keeping in mind that in A and B the spacing and type size was the same (8 point), and that the squares column was next to the No. column in both, the lack of difference in finding squares suggests that style of type face is unimportant (Modern vs. Old Style). The significant difference in favor of A over B in finding square roots is undoubtedly due to the arrangement of columns. In A, the square roots were next to the squares (third column), but in B they were in the sixth column.

In Tables C and E, the type size, spacing down rows, arrangement of columns, and spacing between columns were alike or very similar. This probably explains the lack of difference in finding squares and in finding cube roots.

Lack of difference between D and E in finding square roots and cube roots seems to be due to the fact that the typography is similar in both: thin paper, intercolumnar spacing and rules, 6 point type, and arrangement of columns. Reason for the slight superiority of E over D in finding squares is not clear. Perhaps column arrangement was a factor, for the squares were in the third column in D.

The factors which favor readability of mathematical tables, when the reader starts at the beginning of each table may be summed up as follows: Improved readability is achieved by larger type (8 vs. 6 point), by using at least 1 pica space between columns with no rules (rules probably do not lessen readability provided there is adequate space between columns, e.g., Table B), by separating items down a column into groups of five or ten by leading equivalent to the type size used, by a favorable arrangement of columns across page (as No., square, square root, cube, cube root when one is interested mainly in squares, cubes and roots), and by use of mat white paper thick enough to prevent shadows showing through from print on the reverse side.

When the reader began at page one in finding squares and square roots, as discussed above, the order of the mathematical tables from most to least readable was found to be A, B, C, E, D. Table A was by far the best (errors as well as time) and D was by far the least readable.

When we consider readability uncomplicated by turning pages (given page), the picture is similar but not the same. Tables A, B, C and E are much more readable than D. The main differences in typography common to A, B, C and E in contrast to D are the grouping of items down columns by inserting leading after every fifth or tenth entry, and more adequate spacing between columns. In addition it should be noted that A and B have 8 point type and are printed on thick paper in contrast to D which is in 6 point type on thin paper. The slight superiority of B and C over A is difficult to understand. Table B is in 8, and C in 6 point type while A is in 8 point. The only typographical difference common to B and C in contrast to A is that



in B and C there is only one set of columns present (50 No. items) per page while in A there are three sets of columns of 50 items each in the No., square, and square root columns, i.e., 150 successive items per page. It is possible that the need to locate the proper No. column as well as the numeral hindered the reader somewhat in Table A. It is unlikely that the Old Style numerals in B vs. a modern style in A was important although this should be noted.

The lack of any difference between A and E must have a similar explanation. The advantage of 8 point type in A vs. 6 point in E seems to be nullified by the need to identify the correct No. column as well as the desired numeral in A. The reason for lack of any difference between B (8 point) and C (6 point) is not clear. Two typographical differences should be noted: 1. In C the columns are separated by a 1 pica space while in B they are separated by a 1 pica space plus a rule. 2. In C the column entries are grouped in tens while in B they are grouped by fives. Grouping by tens may have an advantage. It is possible that the intercolumnar spacing and the grouping by tens in C offset the advantage of the larger type in B.

The slight superiority of B over E must be in the quality of the paper (thick vs. thin) and size of type (8 vs. 6). And the slight superiority of C over E may be due to quality of paper plus perhaps the space rather than rules between columns.

There are fewer significant differences between the mathematical tables for finding square roots with page given. A, B, and C are better than D. The essential typographical differences between D and the other tables are lack of grouping down columns, and less adequate spacing between columns and less perhaps quality of paper.

The pattern of differences when looking up cube roots is similar to that for square roots. B and E are better than D. Other differences are not important. Apparently the same typographical factors are operating as in looking up square roots.

The above discussed factors which favor readability of mathematical tables when the reader is given the page on which the num-

ber appears (no turning of pages) may be summarized as follows: grouping of items down columns by inserting ample leading (grouping by tens may be better than by fives), use of one set of columns per page, use of at least a 1 pica space between columns without rules, use of white mat paper thick enough so that shadows from print on the reverse side do not show through, plus perhaps size of type. Variation in style of type face seems unimportant. Apparently the suggestions by Milne (3), Tinker (6) and others (1, 4) that Old Style numerals should be more legible than a modern face when used in tables do not hold.

When the page on which a number was to appear was given the reader in finding squares and square roots, the order of the mathematical tables from most to least readable was (roughly) B, C, A, E, D. The difference between B and C was small.

This experiment was designed as a preliminary investigation of the readability of mathematical tables. Obviously, the experimental design is imperfect. There are too many variables involved. One variable at a time should be studied, or a design should be employed that permits isolation of the variance due to each variable. Nevertheless, the present results suggest that the more important typographical factors favoring good readability are use of at least 8 point type, a favorable arrangement of columns, at least 1 pica space between columns without rules, ample leading to separate entries down columns into groups of five or ten, and paper thick enough to prevent shadows showing through from print on the reverse side of page.

### Summary

1. The purpose of this experiment is to investigate the influence of certain typographical variations upon the readability of mathematical tables.

2. Times in seconds to look up squares, square roots, and cube roots were obtained: (a) when subjects always started with page one of the tables; and (b) when the page containing the number sought was given.

3. Twenty adult subjects served for each of the six parts to the experiment, 120 in all.

4. Five mathematical tables representing a wide range of typographical variations were used.

5. The results of this experiment revealed certain typographical factors which promote more effective readability as well as certain conditions that should be avoided. The evidence educed here suggests the following as an effective typographical arrangement for mathematical tables:

a. Do not crowd an excessive number of columns into a table. This is apt to occur in general purpose tables which include such things as reciprocals, areas, etc. in addition to squares, cubes and roots.

b. Use only one set of columns per page with about 50 entries per column.

c. Use at least 8 point type, either Old Style or a modern face.

d. Employ generous leading to separate numerals into groups of five down columns, and then show grouping into tens by an underline below each tenth row or by some other technique which will be easily noted.

e. Use at least 1 pica space between columns without rules.

f. Use bold face printing in the No. column.

g. Use paper thick enough so that shadows from print on the reverse side of the page do not show through.

h. Use mat white paper and jet black ink to assure maximum contrast between ink and paper.

Received December 8, 1953.

### References

1. Babbage, C. *Tables of logarithms* (Introduction). London: J. Mawman and Co., 1827.
2. Baird, J. W. The legibility of a telephone directory. *J. appl. Psychol.*, 1917, 1, 30-37.
3. Milne, J. R. *The arrangement of mathematical tables*. Ed. by C. Knott, Napier Tercentenary Memorial Volume. London, 1915, 293-316.
4. Report of the committee appointed to select the best faces of type and modes of display for government printing. London: H. M. Stationery Office, 1922.
5. Scott, W. D. *The theory of advertising*. Boston: Small, Maynard and Co., 1903, 119-129.
6. Tinker, M. A. The relative legibility of Modern and Old Style numerals. *J. exper. Psychol.*, 1930, 13, 453-461.

## Comparative and Single Stimulus Methods in Determining Taste Preferences<sup>1</sup>

James A. Bayton and Charles M. Thomas

*Howard University*

Research designed to ascertain relative preferences for variations of a food product is essentially research on the judgmental process. Because of this the investigator must draw upon psychophysics in deriving his methodology (4). The psychophysical methods available fall into two general categories—the methods of comparative judgment and the method of single stimulus (sometimes referred to as the method of absolute judgments). All methods of comparative judgment are alike in that they require the Ss to make direct comparisons of the items in one session. In the method of single stimulus the Ss judge an item without a specific comparison stimulus being present.

In food preference research using the method of comparative judgment two variations are frequently employed—paired comparisons and rank order. When the number of items being investigated is three, the method of paired comparisons has proved efficient (1). With four items the number of pairings is too great for a taste test confined to one session per S; in this instance the method of rank order has been used (2, 3). It can be argued, however, that any comparative judgment procedure is not realistic from the standpoint of actual consumer behavior. How often does the consumer make comparative judgments of the type involved in paired comparisons or rank order experimental designs? The consumer's situation, in which he uses the product without a comparison item present, is a duplication of the method of single stimulus. This being so, a critical question becomes—what is the nature of the ordering of food items, with respect to preference, by the two general procedures? The present research is directed toward this question.

<sup>1</sup> The authors are indebted to Dr. Forrest E. Clements, of the Bureau of Agricultural Economics, for his assistance in this research. The canned orange juices were provided by the Florida Experiment Station and the Bureau of Agricultural Economics.

The food items used were four canned orange juices that varied in °Brix and in Brix-acid ratio. °Brix is a measure of the specific gravity of sugar solutions; the Brix-acid ratio is a measure of the relation between the °Brix and the amount of acid in the solution. Changes in °Brix or Brix-acid ratio are correlated with changes in the tart-sweet quality of the orange juice; the higher the °Brix or Brix-acid ratio the sweeter the juices taste.

### Experimental Design

*Subjects and Materials.* The Ss were 120 individuals 17 years of age and over (68 men; 52 women). The Ss were randomly assigned to the various experimental groups. Each S was tested individually.

The four canned orange juices used were: I. 13.2 °Brix; 18.3 Brix-acid ratio; II. 13.4 °Brix; 9.9 Brix-acid ratio; III. 9.0 °Brix; 18.3 Brix-acid ratio; and IV. 9.3 °Brix; 9.9 Brix-acid ratio.

Variables such as variety of orange and peel-oil content were constant.

The juices were kept under refrigeration so that they were always chilled when used in a test. The juices were served in non-waxed paper cups.

*Procedure.* One-half of the Ss first judged the four juices using the method of rank order and, after a drink of water and a rest period, rated one of the juices under single stimulus conditions. The other half of the Ss reversed this procedure. The particular juice judged under single stimulus conditions was assigned to the Ss in a random manner—one-fourth of the Ss judging a given juice in this particular manner.

The procedure for the method of rank order was as follows. The four juices were placed in a row in front of the S. The original position of the juices varied randomly from S to S. The S tasted the juice on the extreme left and placed it in front of the other three. The juice now on the left in the original row was tasted and placed to the left or right of the first one in terms of, "I like this one better" (placed to right) or, "I like that one



Table 1

Rank Order Preferences for Four Canned Orange Juices

Rank order	13.2 °Brix 18.3 Brix- acid ratio	13.4 °Brix 9.9 Brix- acid ratio	9.0 °Brix 18.3 Brix- acid ratio	9.3 °Brix 9.9 Brix- acid ratio
1	83	25	7	5
2	18	52	34	16
3	13	24	39	44
4	6	19	40	55
Mn	1.52	2.31	2.93	3.24
N	120	120	120	120

$$\chi^2 = 228.36; \text{d.f.} = 9; P < .01.$$

better" (placed to left). Each of the remaining juices was tasted and placed in the new row being developed. When the new order had been established the *S* took a sip of water and tasted the sequence again to verify his order of preference. He was permitted to change the order if he wished. The scoring was 1 to 4, starting with the most preferred juice—the one occupying the extreme right position.

A rating scale was used for the method of single stimulus judgments. The scale was called a "Taste Thermometer." The values ranged from 0 to 100 with gradations of five indicated and the tens numbered. Opposite

100 was the statement, "The best I have ever tasted"; opposite zero was, "The worst I have ever tasted." At 50 was, "Fair; average." The space between 50 and 100 contained the statement, "Better than average" and between 0 and 50 the statement, "Poorer than average." The *S* was told that he would be given one juice to drink and that he would then give it a score. The *Ss* were not told that the juice was one of the four in the rank order procedure.

### Results

The results with the rank order procedure are presented in Table 1. The mean rank for each juice is given but chi-square was used as the test of significance. The order of preference was 13.2 °Brix; 18.3 Brix-acid ratio, 13.4 °Brix; 9.9 Brix-acid ratio, 9.0 °Brix; 18.3 Brix-acid ratio, and 9.3 °Brix; 9.9 Brix-acid ratio. Chi-square tests of all pairings revealed that each juice was significantly different from the other juices.

Table 2 gives the data for the rank order procedure as a function of time of presentation (before or after the single stimulus procedure). In no instance were the rank order distributions for a juice significantly different between sessions.

The analysis of variance results for the single stimulus data are shown in Table 3.

Table 2

Rank Order Preferences for Four Canned Orange Juices by Time of Presentation

Rank Order	13.2 °Brix 18.3 Brix-acid ratio		13.4 °Brix 9.9 Brix-acid ratio		9.0 °Brix 18.3 Brix-acid ratio		9.3 °Brix 9.9 Brix-acid ratio	
	Before single stimulus	After single stimulus	Before single stimulus	After single stimulus	Before single stimulus	After single stimulus	Before single stimulus	After single stimulus
1	38	45	17	8	2	5	3	2
2	11	7	24	28	16	18	9	7
3	8	5	10	14	19	20	23	21
4	3	3	9	10	23	17	25	30
Mn	1.60	1.43	2.18	2.30	3.05	2.82	3.17	3.32
N	60	60	60	60	60	60	60	60
	$\chi^2 = 2.17$		$\chi^2 = 4.65$		$\chi^2 = 2.32$		$\chi^2 = 0.995$	
	d.f. = 3		d.f. = 3		d.f. = 3		d.f. = 3	
	P > .05		P > .05		P > .05		P > .05	

Table 3  
Analysis of Variance for Single Stimulus Ratings of Four Canned Orange Juices

Source of Variation	Sum of Squares	d.f.	Mean Square	F
Juices	4,895.00	3	1,631.67	4.669 (P < .01)
Time of presentation	2,803.34	1	2,803.34	8.022 (P < .01)
Juices × Time of presentation	1,028.34	3	342.78	
Within	39,136.99	112	349.44	
Total	47,863.67	119		

F for variance between juices was significant. Inspection of the mean ratings, however, revealed the following: 13.2 °Brix; 18.3 Brix-acid ratio and 13.4 °Brix; 9.9 Brix-acid ratio had means of 58.00 and 57.83, respectively. The means for 9.0 °Brix; 18.3 Brix-acid ratio and 9.3 °Brix; 9.9 Brix-acid ratio were 46.00 and 43.50, respectively. The differences were significant only when °Brix varied. In other words, the two high °Brix juices were significantly different from the two low °Brix juices. Variation in Brix-acid ratio did not yield significant differences in preference.

The variance between presentations was significant. The mean rating for *all* juices, when the single stimulus procedure came first, was 46.50. The mean rating for *all* juices, when this procedure followed the rank order method, was 56.17.

### Discussion

The above results show that the preference pattern for four canned orange juices is a function of the experimental design used. When the Ss followed the rank order design both °Brix and Brix-acid ratio contributed to preference differentiation. With each S making only a single stimulus rating of one juice (the four juices being randomly assigned to four such groups) preference differentiation occurred only in terms of °Brix. It will be noted that in both methods preference was associated with the relatively higher °Brix (the sweeter juices).

A frame of reference factor was found in the analysis of variance of the data obtained by the method of single stimulus. Starting "cold" with this procedure produced rather low ratings for all juices. When this method followed the rank order procedure the mean

ratings of the juices were appreciably higher.

One limitation that must be placed upon these results rests in the fact that each S did not have experience with the four juices under single stimulus conditions. Another experiment, just completed, indicates that when such is the case no significant differences in mean ratings are obtained for juices that vary in Brix-acid ratio with °Brix held constant. This is the finding in the present experiment.

### Summary

1. Preferences for four canned orange juices that varied in °Brix and in Brix-acid ratio were obtained by a method of comparative judgment procedure (rank order) and the method of single stimulus (using a rating scale).

2. The rank order procedure produced preference differences in terms of °Brix and of Brix-acid ratio.

3. The single stimulus procedure produced differences only in terms of °Brix.

4. From both methods it appears that preference is associated with juices of relatively higher °Brix (the sweeter juices).

Received November 20, 1953.

### References

1. Bayton, J. A. Consumer preferences for selected frozen concentrated apple juice. *Bureau of Agricultural Economics*, June, 1951.
2. Bayton, J. A. and Bell, H. P. Discrimination tests and preliminary preference ratings of frozen concentrates for lemonade. *Bureau of Agricultural Economics*, September, 1952.
3. Bell, H. P. and Bayton, J. A. Taste tests on canned orange juice. *Bureau of Agricultural Economics*, June, 1953.
4. Clements, F. E. Psychophysical methods in market research. *Florida State Horticultural Society*, 1951, 64, 148-153.

## Method of Single Stimulus Determinations of Taste Preference<sup>1</sup>

Forrest E. Clements, James A. Bayton, and Hugh P. Bell

*United States Department of Agriculture, Washington, D. C.*

There are two fundamental considerations that would lead one to select a method of single stimulus approach in research on preferences for variations of a food product. First, the method of single stimulus is a duplication of the situation that is typical for consumers. Seldom does the consumer have available in his home at a given time several variations of a particular food product—the situation that would be conducive to making preference judgments based upon the direct comparisons involved in the method of comparative judgments. Realistic research on taste preference should attempt to utilize the actual home situation.

The second factor that will force the experimental design into a method of single stimulus model is the number of the variations of the food product being tested since adaptation is a variable. It has been our experience, in either laboratory or home situations, that three items is the maximum number for a paired comparison design when testing occurs in one session; with a rank order design four items is the maximum (1, 2, 3, 4, 5). When the number of items is five or more comparative judgment models should be abandoned for a method of single stimulus design.

Of necessity, any method of single stimulus design for determining taste preference will require the use of some type of rating scale or scoring system. Most scales used in such research are either point-scales with terse definitions of each point or "thermometers" allowing for 0 to 100 scoring with descriptions such as "Excellent," "Good," etc., at selected points. When designing a research project that will involve a large-scale sample of a cross-section of consumers the scale or scor-

ing system used will have to be easily understood. Because of this, a third type of scale was investigated in the present experiment. This was a highly unstructured scale with only the extremes of the continuum defined. None of the points available for choice as expressing degree of preference was identified or defined. The primary purpose of this experiment was to test the relative efficiency of the three types of scales in determining taste preferences when the research is conducted with a method of single stimulus design under realistic conditions.

The particular food items used were three canned orange juices that varied in Brix-acid ratio with °Brix held constant. °Brix is a measure of the specific gravity of sugar solutions; Brix-acid ratio is a measure of the relation between the °Brix and the amount of acid in the solution. The lower Brix-acid ratios are characterized by tart-sour taste; the higher Brix-acid ratios are sweeter in taste. In addition, when °Brix is constant and amount of acid varies the juices change in body or consistency; the higher Brix-acid ratios tend to be "thinner" as well as sweeter.

This research was preliminary to a large-scale investigation involving six canned orange juices. To facilitate this preliminary project the lowest and highest Brix-acid ratios in the six juices and one from the middle set were used.

### Procedure

**Scales.** Scale A ranged from 0 to 100 with gradations of five numbered. To the right of the scale certain areas were bracketed and labeled. The area 90-100 was "Excellent"; 70-90 was "Very Good"; 50-70 was "Good"; 30-50 was "Fair"; 10-30 was "Poor"; 0-10 was "Very Poor." The Ss were instructed first to decide what they thought of the juice in a general way—"Very Good," "Poor," etc.—and then to rate it by assigning a score in the particular area.

Scale B consisted of ten 5/16" squares arranged vertically. Above the top square was "Excellent"; below the bottom one was "Very Poor." No other definitions or descriptive state-

<sup>1</sup>The canned orange juices used in this experiment were furnished by the Florida Experiment Station. The authors wish to acknowledge the efforts of Mrs. Motier Fisher and Mrs. Mary George Robinson who did the necessary field work. Discussions with Dr. Franklin R. Kilpatrick led to our development of Scale B.



ments were given. The Ss were shown that their opinion about the juice could be expressed as falling anywhere from "Very Poor" up through "Excellent." They were to put a cross in the square that expressed their opinion. Scoring for Scale B was 1 to 10, starting with the bottom square.

Scale C was the following 7-point scale:

Excellent—the best *canned orange juice* I have ever tasted.

Good—much better than *other canned orange juice* I have tasted; but not the best.

Fair—a little better than *other canned orange juice* I have tasted; but not much better.

Borderline—can't decide whether it is better or worse than *other canned orange juice* I have tasted.

Poor—a little worse than *other canned orange juice* I have tasted; but not much worse.

Very Poor—much worse than *other canned orange juice* I have tasted; but not the worst.

Objectionable—the worst *canned orange juice* I have ever tasted.

The Ss were instructed to check the square preceding the statement that expressed their opinion about the juice. The scoring was 1 to 7, starting with "Objectionable."

*Descriptive Check-List.* After rating a juice the Ss were asked to check those items on a list that they thought described it. They could check as many items as they thought applied. The check-list contained items such as, "Too sweet," "Too tart or sour," "Just the right sweetness," etc.

*Experimental Design.* The Ss were adult members of households in a new residential area adjacent to Alexandria, Virginia. The area consisted of about 600 homes. Approximately every seventh home was contacted, yielding a panel of 90 households. To be eligible a household had to contain at least two adults who agreed to participate throughout the experiment.

Test I. The purpose of Test I was to obtain preference ratings for the three canned orange juices on each of the three scales. The 90 households were divided into three sets of 30 each. Each set received a given scale. On the first placement of the juices 10 households using a given scale received 12 Brix-acid ratio, 10 received 16 Brix-acid ratio, and 10 received 22 Brix-acid ratio. These assignments were made in a random manner. Three placements were made per household until each S had experience with all of the juices. Three to four days intervened between placements. The homemakers were instructed to place the juice in the refrigerator overnight and to make the tests the following day.

The juices were in unlabeled 10-ounce cans, coded for identification purposes. Only one can of juice was left at a household per placement.

This was done so that all the juice would be consumed when tested. As stated, at least two Ss per household made the tests.

Test II. The purpose of Test II was to investigate the reproducibility of the ratings for the 12 Brix-acid ratio juice on each scale. Fifteen of the 30 households that worked with a given juice in Test I were selected. The Ss were told merely that they were rating "another" juice in our set. This test was conducted about one month after the completion of Test I.

Test III. The purpose of Test III was to investigate the reproducibility of the preference relationship obtained with Scale B in Test I for the 12 and 22 Brix-acid ratio juices. This test took place approximately two months after Test I. The 30 households that had worked with Scale B in Test I took part in this test. One-half of the households received the 12 Brix-acid ratio juice on the first placement, the remainder received 22 Brix-acid ratio on the second placement; after three days, the juices were reversed.

## Results

The mean preference ratings on Scale A for the 12, 16, and 22 Brix-acid ratio juices were 61.0, 60.1, and 57.4, respectively. The differences were not significant. Scale B yielded mean ratings for the respective juices of 5.7, 5.8, and 5.8. On Scale C the means were 5.3, 5.1, and 5.0. Neither of the latter two scales produced significant differences between the juices.

Table I presents the preference data in terms of whether an S scored the 12 Brix-acid ratio juice above or below the mean for that juice on a given scale. Those who scored the juice above the mean were designated as the "Like" group; those scoring it below the mean were called the "Dislike" group. On each scale, the "Like" 12 Brix-acid ratio group gave a significantly higher rating to that juice than to either the 16 or the 22 Brix-acid ratio juices. In these groups, however, the ratings for the latter two juices were not significantly different. Conversely, on each scale, the "Dislike" 12 Brix-acid ratio group tended to give higher ratings to the 16 and 22 Brix-acid ratio juices than to the 12 Brix-acid ratio juice. These differences were not significant for Scale A. In Scale C, the difference between 12 and 16 Brix-acid ratio was significant; the difference between 12 and 22 Brix-acid ratio was not significant. Both of the differences involved were significant on

Table 1  
Preference Scores for Canned Orange Juices that Vary in Brix-acid Ratio by  
"Liking" vs. "Disliking" Brix-acid Ratio 12 (Test I)

	Brix-acid ratio (12 °Brix)			<i>t</i> for mean difference		
	12	16	22	12 vs. 16	12 vs. 22	16 vs. 22
<b>"Like" Brix-acid Ratio 12</b>						
Scale A (N = 30)						
Mn	76.7	66.8	61.3	3.2**	4.1**	1.3
SD	7.0	16.0	17.7			
Scale B (N = 32)						
Mn	7.5	5.5	5.7	4.0**	3.5**	0.3
SD	1.1	2.5	2.6			
Scale C (N = 35)						
Mn	6.3	5.2	5.2	5.0**	4.4	0.2
SD	0.5	1.2	1.5			
<b>"Dislike" Brix-acid Ratio 12</b>						
Scale A (N = 29)						
Mn	44.8	53.2	53.3	2.0	1.8	0.1
SD	11.4	18.5	19.3			
Scale B (N = 25)						
Mn	3.4	6.2	6.1	5.4**	4.1**	0.4
SD	1.4	1.9	2.4			
Scale C (N = 25)						
Mn	4.0	5.0	4.6	3.2**	1.3	0.8
SD	1.2	1.2	1.8			

\*\* Significant at the 1 per cent level.

Scale B. For the "Dislike" 12 Brix-acid ratio group, none of the 16-22 Brix-acid ratio differences were significant.

A similar analysis was made with the 16 Brix-acid ratio juice. The general pattern was that when this particular juice was "liked" its mean was higher than those for either the 12 or 22 Brix-acid ratio juices. The respective differences involved were significant on each scale. When the 16 Brix-acid ratio juice was "disliked" its mean was lower than those for the 12 or 22 Brix-acid ratios. Both of the differences involved were significant only for Scales B and C.

Table 2 repeats the above analysis in terms of "Like" and "Dislike" 22 Brix-acid ratio for each scale. The pattern in this instance was for those who "liked" the 22 Brix-acid ratio juice to give the other two juices lower scores. Both of the particular differences involved were significant on Scales A and B. Those who "disliked" the 22 Brix-acid ratio juice gave the other two juices

higher scores than observed for the 22 Brix-acid ratio. Both differences were significant on each scale.

The data on reproducibility of the preference ratings for the 12 Brix-acid ratio juice, after a month had passed, showed that for each scale the difference in preference rating was not significant.

The results on reproducibility of the preference data on Scale B for the 12 and 22 Brix-acid ratio juices two months after Test I demonstrated that the mean preference ratings for the two juices again were not significantly different. However, the division of the Ss into "liking" and "disliking" a respective juice revealed a pattern similar to that obtained in the original analysis. Those who "liked" a given juice tended to give the other one lower ratings; when a juice was "disliked" the other juice was given higher ratings. The difference was not significant, however, for those who "disliked" the 12 Brix-acid ratio juice.

Table 2

Preference Scores for Canned Orange Juices that Vary in Brix-acid Ratio by  
"Liking" vs. "Disliking" Brix-acid Ratio 22 (Test I)

	Brix-acid ratio (12 °Brix)			<i>t</i> for mean difference		
	12	16	22	12 vs. 16	12 vs. 22	16 vs. 22
<b>"Like" Brix-acid Ratio 22</b>						
Scale A (N = 30)						
Mn	62.2	64.3	73.0	0.5	2.9**	2.4*
SD	18.7	18.8	10.5			
Scale B (N = 32)						
Mn	5.4	6.4	7.8	1.5	4.4**	3.8**
SD	2.5	2.1	1.3			
Scale C (N = 44)						
Mn	5.5	4.9	5.8	1.8	1.2	3.9**
SD	1.5	1.3	0.8			
<b>"Dislike" Brix-acid Ratio 22</b>						
Scale A (N = 29)						
Mn	59.8	55.9	41.2	1.0	4.6**	3.5**
SD	19.2	18.9	10.7			
Scale B (N = 25)						
Mn	6.1	5.0	3.4	1.7	5.3**	3.0**
SD	2.2	2.4	1.2			
Scale C (N = 16)						
Mn	4.9	5.4	2.6	1.7	5.5**	8.8**
SD	1.3	1.0	1.1			

\* Significant at the 5 per cent level.

\*\* Significant at the 1 per cent level.

Table 3

Descriptions of Canned Orange Juices that  
Vary in Brix-acid Ratio

Description	Brix-acid ratio (12 °Brix)		
	12	16	22
	Per cent	Per cent	Per cent
Too tart or sour	39	15	11
Too sweet	11	21	28
Too thin or watery	22	27	35
Too artificial	25	31	30
Just the right sweetness	35	44	38
Just the right tartness or sourness	30	18	17
Does not taste like fresh orange juice, but still is pretty good	51	55	53
Tastes like fresh orange juice	11	9	7
Number	175	175	175

Note: Percentages add to more than 100 because  
some Ss checked more than one descriptive item.

The descriptions of the three canned orange juices by all Ss, regardless of scale used, are presented in Table 3. The percentage of Ss who described the juices as being too tart or sour decreased from the 12 to the 22 Brix-acid ratio. The percentage of Ss describing a juice as too sweet increased from the 12 to the 22 Brix-acid ratio. "Too thin or watery" was most frequently given for the 22 Brix-acid ratio juice. Approximately 50 per cent of the Ss said that although these juices did not taste like fresh orange juice they still were "pretty good."

In Table 4 the descriptions have been analyzed in terms of those "liking" or "disliking" the 12 and 22 Brix-acid ratio juices. Those who "disliked" the 12 Brix-acid ratio juice tended to describe it as too tart or sour and too artificial. The Ss who "disliked" the 22 Brix-acid ratio juice tended to say it was too sweet, too thin or watery, and too artificial. Approximately 50 per cent of those who liked



Table 4

Descriptions of Canned Orange Juices by "Liking" vs. "Disliking" Brix-acid Ratio 12 and Brix-acid Ratio 22 (12 °Brix)

Description	12 Brix-acid ratio		22 Brix-acid ratio	
	"Like"	"Dislike"	"Like"	"Dislike"
	Per cent	Per cent	Per cent	Per cent
Too tart or sour	23	56	8	14
Too sweet	10	10	18	38
Too thin or watery	15	28	23	47
Too artificial	8	46	10	53
Just the right sweetness	51	10	57	17
Just the right tartness or sourness	44	10	28	1
Does not taste like fresh orange juice, but still is pretty good	74	28	74	30
Tastes like fresh orange juice	17	1	13	1
Number	100	78	100	81

Note: Percentages add to more than 100 because some Ss checked more than one descriptive item.

a juice said it had "just the right sweetness." "Just the right tartness or sourness" was more frequently used to describe the 12 and the 22 Brix-acid ratio among those who "liked" these respective juices.

### Discussion

Test I can be viewed as three independent experiments on preferences for these canned orange juices; each experiment involving a different scale. The general pattern of the preference results was similar for each scale. Regardless of the scale, the means of the preference ratings, for all Ss, were not significantly different. It had been expected that the 12 Brix-acid ratio would be too tart and the 22 Brix-acid ratio too sweet, thus producing the highest mean preference ratings for the 16 Brix-acid ratio juice. This expectation was based upon the assumption that we would be dealing with a sample from one population. Taking the means for all Ss per juice at their face value one would conclude that one juice was as likely to be preferred as another. This, in turn, would raise the question of whether the Ss really could distinguish between the three juices in this method of single stimulus approach although prior comparative judgment experiments have shown that these juices are discriminable (3).

The division of the Ss into "liking" or "disliking" a given juice gave evidence that, with respect to canned orange juices, there are two basic populations that we sampled. One population likes a tart juice; the other likes a sweet juice. Those who "liked" a tart juice gave it a relatively high score and gave a low score to the sweeter juice. The Ss who "disliked" the tart juice gave it a low score and assigned a relatively high score to the sweeter juice. Obviously, this phenomenon produced a cancelling-out effect on the means per juice for all Ss. This effect is particularly striking since the results were obtained with a method of single stimulus experimental design.

That this phenomenon is no artifact is seen in its demonstration in Test I with three different scales. Further evidence is seen in the replication of the experiment (Test III) after two months, using the 12 and 22 Brix-acid ratio juices. Once again, the difference between the juices, for all Ss, was not significant. However, analysis in terms of "liking" and "disliking" showed that the Ss who "liked" one juice scored it relatively high in contrast to the score given the other juice.

The data from the descriptive check-list show that the Ss were responding to the critical variables in these juices (tartness-sweet-

ness and body or consistency). Furthermore, the data from all Ss yield additional support for the conclusion that two populations were involved. The percentage describing a juice as too tart or sour decreased from the 12 through 22 Brix-acid ratio juices. The reverse was true for those calling these juices too sweet. When the descriptions were considered in terms of "liking" or "disliking" one of these juices they revealed that the Ss were responding to the tart-sweet dichotomy.

The Ss were asked whether they liked orange juice "somewhat on the tart side or somewhat on the sweet side." Forty-seven per cent said they liked it tart, 46 per cent replied "sweet," and 7 per cent volunteered the information that they liked it "medium," "in-between," etc. Although the distribution of these replies again supports the two-population concept they were not indicative of how the juices were scored. There was only low correlation between the replies to this question and the preference scores for the juices. This can only mean that the question does not locate those Ss who actually prefer tartness or sweetness under direct experience with the juices.

The question now becomes whether there was any difference in the efficiency of the three scales in revealing the preference pattern. It has been pointed out that the preference pattern was similar for the three scales. Inspection of the *t*'s for the "Like"- "Dislike" data shows that Scale B tended to give higher significance values than the other two scales. The median *t* for Scale A was 2.00, for Scale B was 3.04, and for Scale C was 2.51.

The reproducibility test with the 12 Brix-acid ratio juice did not produce significantly different ratings on any scale. However, it should be noted that Scale B came closer to doing this than did the other two scales.

There is an indication that as the Ss continued to work with these juices the preference ratings tended to rise. In Test I the means for all Ss for the 12 and 22 Brix-acid ratio were 5.70 and 5.84, respectively. In Test III the respective means were 6.73 and 6.17. This supports the prior finding that

repeated experience with the juices, under single stimulus conditions, produces generally higher ratings (5). In spite of this general increase in preference the "Like"- "Dislike" patterns still existed.

On the basis of the results of this experiment it was decided to use Scale B in a 720 household study of preferences for six canned orange juices that vary in Brix-acid ratio, with °Brix constant. Scale B seems to be somewhat more efficient in revealing preference patterns and has the advantage of minimizing language and intellectual difficulties.

### Summary

1. Using a method of single stimulus design three canned orange juices that varied in tartness-sweetness and in body or consistency were given preference ratings. Three different scales were used, each *S* working with only one scale.

2. Under method of single stimulus conditions the Ss were able to respond to the variables of tartness-sweetness and body or consistency.

3. The results indicated that there are two populations with respect to preference for canned orange juice—one prefers a tart juice, the other a sweet one.

4. A relatively unstructured scale, with only the ends of the continuum defined, tended to be most efficient. All three scales, however, revealed the same pattern of preference.

Received December 12, 1953.

### References

1. Bayton, J. A. Consumer preferences for selected frozen concentrated apple juice. *Bur. agri. Econ.*, June, 1951.
2. Bayton, J. A. and Bell, H. P. Discrimination tests and preliminary preference ratings of frozen concentrates for lemonade. *Bur. agri. Econ.*, September, 1952.
3. Bell, H. P. and Bayton, J. A. Taste tests on canned orange juice. *Bur. agri. Econ.*, June, 1953.
4. Clements, F. E. Psychophysical methods in market research. *Fla. hort. Soc.*, 1951, 64, 118-153.
5. Thomas, C. M. *Comparative vs. single stimulus methods in determining taste preferences*. Unpublished master's thesis, Howard Univ., 1952.

# The Effect on Performance of Tilting the Toll-Operator's Keyset \*

Edythe M. Scales

*Bell Telephone Laboratories Inc., Murray Hill, New Jersey*

and

Alphonse Chapanis

*Department of Psychology, The Johns Hopkins University*

Although engineers and human engineers frequently recommend the use of inclined visual displays and panels on control consoles, there is virtually no scientific evidence to show that this design practice has any measurable effect on operator efficiency (2). The present study was undertaken to discover whether tilting the keyset now used by long-distance operators would have any effect on their keying performance. In view of the lack of experimental evidence in this area, we believe that our findings may be of general interest.

## Experimental Method

**Apparatus.** The long-distance operator's keyset is a ten-button set with the numbers and let-

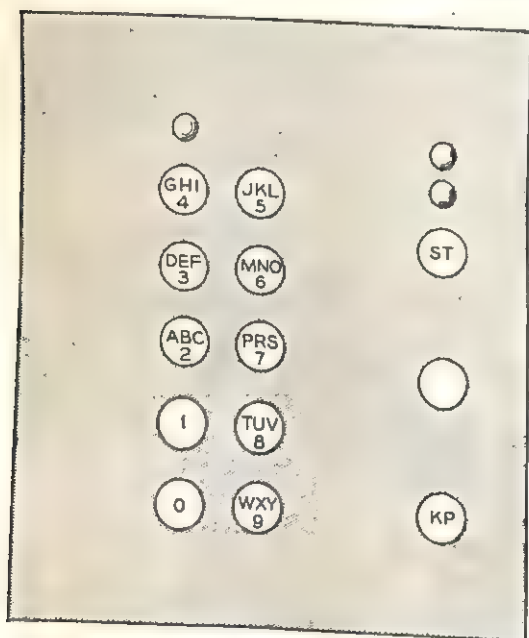


FIG. 1. Top view of a toll-operator's keyset.

\* The experiment reported here was done at the Bell Telephone Laboratories.

ters arranged in two vertical rows of five. A third column contains two keys, the KP key, which sets up the apparatus to receive a number sequence, and the ST key, which clears the machine of the sequence just keyed (see Figure 1). The keyset is normally mounted on a horizontal working surface about 13 inches back from the front edge of this surface, and 9 inches to the right of center. The experimental apparatus shown in Figure 2 approximates the toll-operator's position in its essential dimensions. In normal operation, the keyset is horizontal. In the present study, the keyset was mounted on hinges so that it could be inclined at eight angles relative to the working surface: 0, 5, 10, 15, 20, 25, 30, and 40 degrees.

A remotely-located recorder printed numbers corresponding to the ten number-letter keys of the keyset.

The illumination of the experimental room was constant throughout the test at an adequate intensity.

**Materials.** The stimuli for the keying task were ten-place number and letter combinations of the following form: 3 digits, space, 2 letters 1 digit, space, 4 digits. In long-distance operation the first three digits are the "code" to the distant location, the two letters and the next digit give the subscriber's exchange, and the remaining four digits the subscriber's number. For these tests, the numbers were obtained from a table of random numbers. The letters were also selected randomly from a special table which ensured that all letter combinations appeared equally often (except that the letters Q and Z were never used). The stimuli were presented to the subject in list form. Twenty-four different lists, each containing 50 stimuli, were used in the first part of the experiment (practice sessions); 256 different lists, each containing 100 stimuli, were used in the second part (test sessions).

**Experimental Design.** The experiment was divided into two consecutive parts, practice sessions and test sessions, each part extending over eight days. Two pairs of 8 by 8 Latin squares (four in all) were used, the main-effect variables of each being: (1) subjects; (2) days; and (3) inclinations of the keyset. One pair of identical Latin squares was assigned to the practice ses-



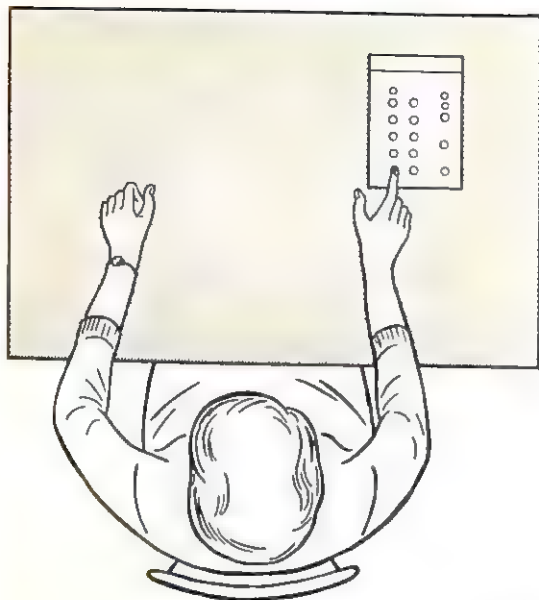


FIG. 2. A schematic illustration of the toll-operator's position in this experiment.

sions; the other pair of identical Latin squares to the test sessions. One square of each pair was assigned to 8 male subjects and the other to 8 female subjects.

**Procedure.** Instructions read to each subject at the beginning of the first practice session covered these essential points:

1. Position of the subject with respect to the keyset.
2. Technique of keying. (All keying was done with the first or second finger of the right hand.)
3. The criterion. (Primary emphasis was placed on accuracy.)
4. Procedure to follow when an error was discovered.

Because this last instruction required the subject to stop and rekey the entire number whenever he thought he made an error, the time and error measurements are not entirely independent. Later on, however, we will see that this is not an important consideration.

On each day of the practice sessions, all subjects keyed three number lists with a 5-minute rest between lists. On each day of the test sessions, all subjects keyed two number lists with a 10-minute rest between lists.

**Subjects.** Sixteen subjects, eight male and eight female, participated in this study. Their ages were between 18 and 35. No subject had previous experience on this keyset. One female subject did not participate on the last day of the test sessions.

## Results

All data expressed in the following graphs have been computed from individual error and time values. An individual error value is the percentage of incorrect keyings made by a subject, based on keying 150 numbers in a practice session or 200 numbers in a test session. An individual time value is the total time required for a subject to key the three number lists in a practice session or the two number lists in a test session.

For this kind of experimental design, an analysis of variance is usually employed to evaluate the data. Although such analyses were carried out in the present study, the results are much more clearly described in the accompanying graphs. All graphs depict three statistical measures: (1) the arithmetic mean; (2) the mean plus and minus one standard deviation; and (3) the total range of values.

**Effect of Tilt.** Figures 3 and 4 clearly

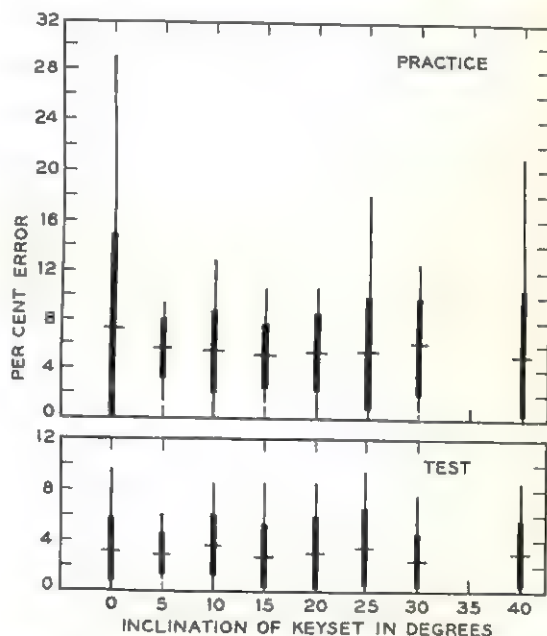


FIG. 3. The data at each inclination are based on the percentage of errors made by 16 subjects each of whom keyed 150 numbers (practice sessions) or 200 numbers (test sessions). (At 25° for the test sessions there were only 15 subjects.) The short horizontal line is the mean, the solid vertical bar the mean plus and minus one standard deviation, the thin vertical bar the range of individual error percentages.

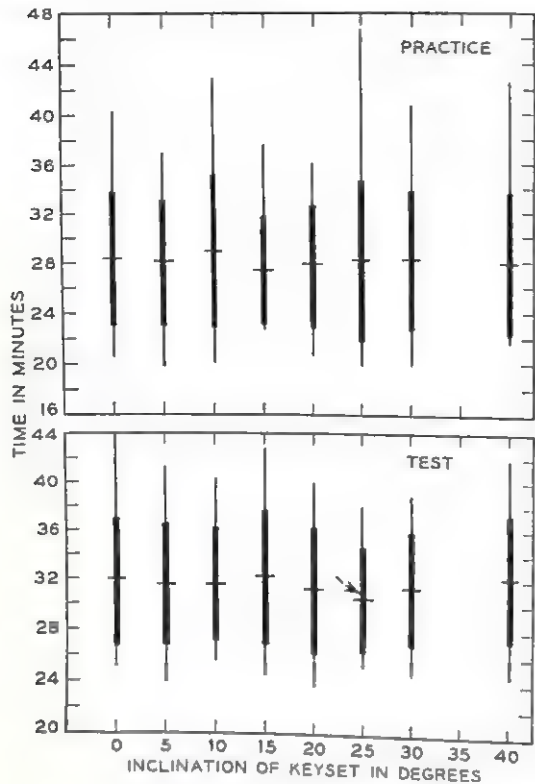


FIG. 4. These time data correspond to the error data of Figure 3. The basic datum is the total time required by each subject to key a set of numbers. The mean, standard deviation, and range are represented as in Figure 3. The arrow at 25° shows the mean estimated for 16 subjects (see text).

demonstrate that inclination of the keyset has virtually no effect on keying performance, either in terms of error or time. Figure 3, for example, shows that the averages for the test sessions are within a small range: 2.5 to 3.7 per cent. Moreover, there is no evidence of any systematic trend in the mean values as a function of keyset inclination. A straight line with zero slope appears to fit these data adequately. It is not likely that the data are appreciably affected by the fact that one subject was not tested at the 25-degree inclination.

Figure 4 also shows that the average times lie within a small range. For the practice sessions this range is 1.5 minutes (27.3 to 28.8 minutes). For the test sessions the range is 1.1 minutes (31.1 to 32.3 minutes) provided that the estimated value for 25 degrees is used. Since the subject who was not

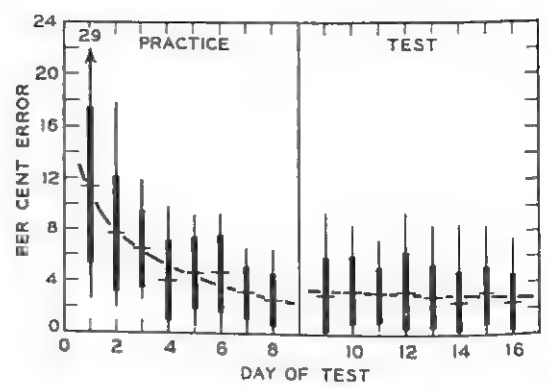


FIG. 5. The error data for each day are based on the performance of 16 subjects, except that on the last day there were only 15 subjects. The mean, standard deviation, and range are represented as in Figure 3. The curves through successive means were drawn by inspection.

tested at 25 degrees (Subject E, Figure 8) had the longest average keying time, the mean for 25 degrees is undoubtedly too low because of this omission. The arrow in Figure 4 shows the estimated value for the mean on the assumption that Subject E had turned in a value equal to her average keying time.

**Learning.** Figures 5 and 6 show the course of learning in terms of errors and time, respectively. Both show a large and significant decrease due to learning throughout the first eight days of test. Errors do not show a significant decline in the second eight days,

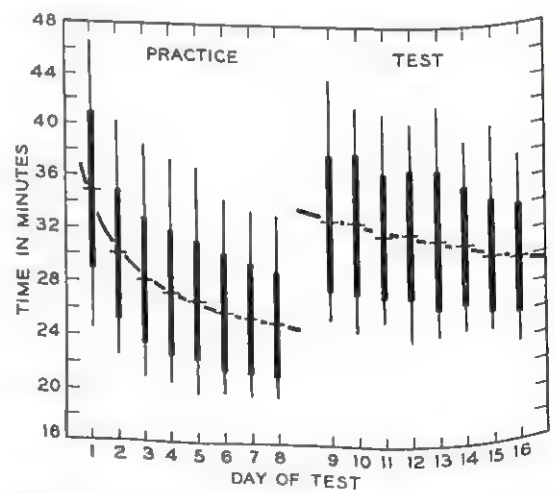


FIG. 6. These time data correspond to the error data in Figure 5. The mean, standard deviation, and range are represented as in Figure 3. The curves through successive means were drawn by inspection.

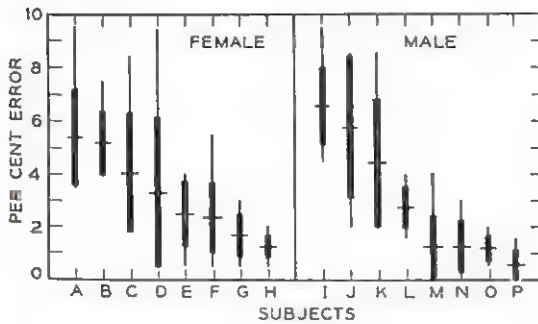


FIG. 7. These error data show each subject's performance for the eight test sessions. (The data for E are based on only seven test sessions.) The mean, standard deviation, and range are represented as in Figure 3.

although the keying times do. It is apparent, therefore, that learning was not complete even after 16 days of test. In this experiment the effects attributable to learning are much greater than those produced by variations in the inclination of the keyset.

**Individual Differences.** Figures 7 and 8 are plots of individual keying performances in terms of error and time for *only* the test sessions. These graphs show the most important source of variance in our experiment. The averages in Figure 7 cover a range from 0.5 to 6.6 per cent. In Figure 8 the range is from 24.8 to 39.7 minutes.

Rank-order correlation coefficients between average errors and average times for the test sessions were computed for the male and female subjects separately. For the female subjects, the coefficient was  $+0.07$ ; for the

male subjects,  $-0.76$ . We have no explanation for the difference between the magnitudes of these two correlation coefficients.

### Discussion

In the specific work situation of the present study, we have found performance to be unaffected by the inclination of the working surface. However, spontaneous comments from all of our subjects indicated that they preferred an inclined keyset surface to a horizontal one. Furthermore, about half of the subjects expressed a preference for a keyset inclination between 15 and 25 degrees.

These subjective preferences, as well as the quantitative data, are in agreement with another specific investigation that was concerned with speed and accuracy of target indication on a radar which was mounted at various inclinations (1). Since the nature of the tasks in these two situations differs so radically, the agreement between the two sets of results suggests that we can perhaps apply the findings to other work situations. If a working surface is clearly visible to the operator and if it is within easy reach, inclining the work surface will probably not result in any measurable effect on performance. People seem to like inclined surfaces better than horizontal ones, but we have no way of evaluating the importance of such preferences.

Many of the standard deviations in Figures 2 through 7 are large because the data are not homogeneous, i.e., they include several sources of variance. For example, the standard deviations for each keyset inclination in Figures 3 and 4 include the differences between subjects and the differences between days, both of which are large. In Figures 5 and 6, the standard deviations include differences between subjects and between inclinations. In this case the standard deviations are smaller because, as we have seen, variations produced by keyset inclination are small. In Figures 7 and 8, the standard deviations are small because the variations attributable to inclinations and days (for the test sessions *only*) are also small.

Earlier we noted that the time and error scores are not independent. If this were an

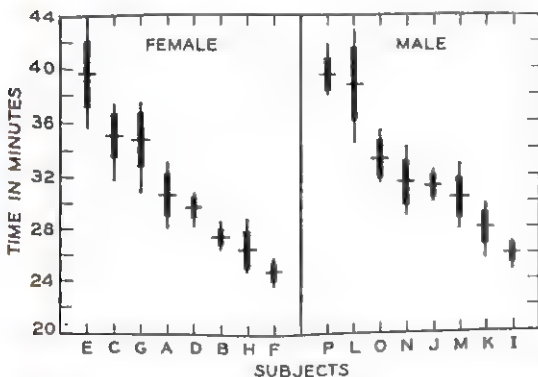


FIG. 8. These time data correspond to the error data in Figure 7. The mean, standard deviation, and range are represented as in Figure 3.



appreciable factor in this experiment, we should expect the two values to be positively correlated. Actually they are not. In addition, we should note that in the test sessions there were only a few errors committed and, of those made, less than one in three was detected and rekeyed by the subject. All in all, therefore, we do not believe that this is an important consideration in these data.

### Summary

The present experiment investigated two measures of keying performance, accuracy and time, as a function of inclination of the keyset. The keyset was inclined at eight angles, 0, 5, 10, 15, 20, 25, 30, and 40 degrees, relative to the working surface.

The test was divided into two parts, practice sessions and test sessions. The subject's task was to key lists of ten-place number and letter combinations. Eight by eight Latin squares were used, the principal variables be-

ing subjects, days, and inclinations of the keyset. The results clearly demonstrate that:

1. Keying accuracy and keying time are independent of the inclination of the keyset.
2. Both accuracy and speed increased significantly throughout the sixteen days of test.
3. The greatest source of variation in this experiment is that produced by differences between subjects.

*Received August 12, 1954.*

*Early publication.*

### References

1. Leyzorek, M. *Mounting angle of a VJ remote radar indicator and its effect on operator performance*. Special Devices Center, Office of Naval Research, Report No. 166-I-41, February 1948.
2. Stellar, E. Human factors in panel design. Chapter 6 in Panel on Psychology and Physiology, Committee on Undersea Warfare. *A survey report on human factors in undersea warfare*. Washington, D. C.: National Research Council, 1949. Pp. 153-175.

## The Use of a Joy-Stick in Making Settings on a Simulated Scope Face \*

William Leroy Jenkins and A. Charles Karr

*Lehigh University*

An earlier study<sup>1</sup> reported the use of levers in making settings on a linear scale. The most important variable proved to be the ratio between the movement of the lever tip (L) and the movement of the pointer (P). An L/P ratio of approximately three was found to be optimal. The current investigation extends the problem into two dimensions, using a joy-stick to set a cursor on a simulated scope face.

An operational diagram of the apparatus is shown in Figure 1. A vertical twelve-inch aluminum disc, with its center at approximately eye-level and about 24" from the subject's eyes, simulates a scope face. Seven quarter-inch circular lucite inserts are spaced around a ten-inch diameter, six inserts around a seven-inch diameter, and four around a three-inch diameter. The cursor (a brass disc .150" in diameter) is controlled by a joy-stick placed between the subject's knees with its tip about six inches below the edge of the simulated scope.

Right-left components of the joy-stick movement are transmitted through the lower shafts to the small pulley and then to the upper shaft, causing the long cylinder to move right and left across the simulated scope face. Various ratios of movement between joy-stick and upper shaft are obtained by shifting the belt attachments along the bar at the end of the upper shaft.

Front-back components of the joy-stick movement operate a hydraulic pump that serves to move the piston up and down in the long cylinder. Ratios between movement of the joy-stick and movement of the piston are changed by sliding the attachment of the hydraulic pump up or down on the joy-stick.

\* This research was executed under Contract AF 18(600)-24 between the Institute of Research, Lehigh University, and the USAF Wright Air Development Center, Aero Medical Laboratory, Wright-Patterson Air Force Base, Dayton, Ohio.

<sup>1</sup> Jenkins, W. L. and Olson, M. W. The use of levers in making settings on a linear scale. *J. appl. Psychol.*, 1952, 36, 269-271. Also USAF Technical Report No. 6563.

Since the viscous friction of the right-left system is less than that of the front-back system, it is necessary to equalize the kinesthetic feel by adding viscous friction to the right-left system. This is done by adjusting a Prony brake, liberally coated with graphite lubricant, which adds a viscous drag, until the right-left viscous friction seems equal to the front-back friction.

The cursor and scoring mechanism are mounted at the top of the piston that moves up and down in the long cylinder. The scoring mechanism operates as follows: When the subject has completed a setting he pushes a switch which discharges a condenser into a small electromagnet. The electromagnet moves the lucite strip bearing the brass cursor, so that the brass disc comes in contact with the scope face for a fraction of a second. If the cursor touches only the lucite insert, no electrical contact is made and a green light glows. If the cursor is not entirely within the confines of the insert, electrical contact is made between the brass disc and the aluminum scope face, lighting a red light to indicate a mis-setting.

### Procedure

The procedure for a single setting is as follows: Following a ready signal, the experimenter moves a switch that simultaneously lights one of the inserts and starts the timing clock. The subject moves the joy-stick to bring the cursor onto the lighted insert, and then pushes a button that simultaneously operates the scoring mechanism and stops the timing clock. The elapsed time on the clock shows the setting time, and a green or a red light indicates whether the setting is correct.

### Results

For clarity, the results will be described in five parts, paralleling the chronological order of the experiments.

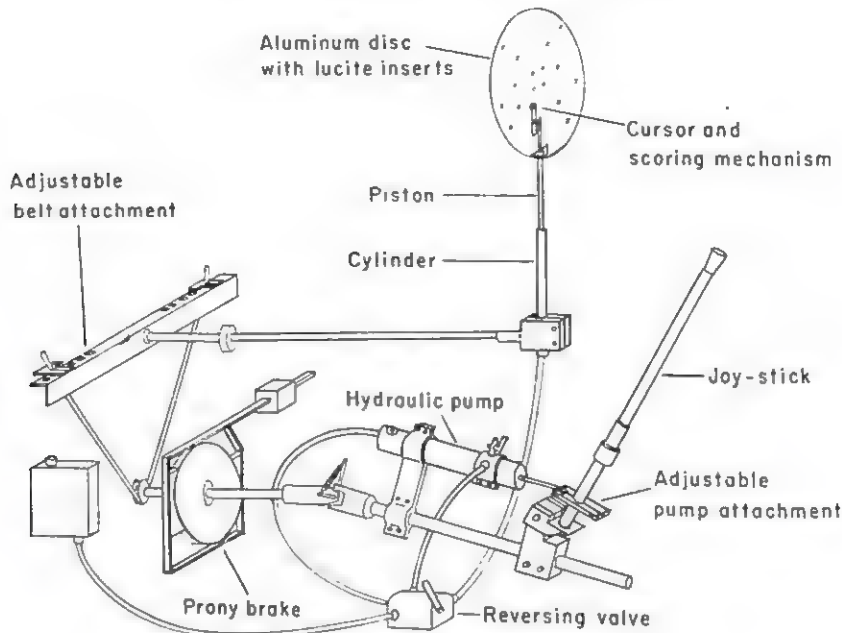


FIG. 1. Operational Diagram of Joy-Stick Apparatus.

*L/P Ratio in General.* By combining various lever lengths and apparatus settings, twelve L/P ratios between 1.0 and 3.9 were tested, with nine target positions. Each of 20 subjects made 20 settings at each combination of L/P ratio and target position.

Table 1 shows for the twelve L/P ratios the mean setting time, variability, and mis-settings. In all three respects, ratios of 2.0 and above appear to be clearly more favorable than the ratios of 1.7 and under. Although the highest L/P ratios are obtained with the longer levers, each of the longer levers is also represented among the lower (unfavorable) L/P ratios.

When the data are re-analyzed to compare the favorable ratios (2.0 and up) with the unfavorable ratios (1.7 and down) for each of the 20 subjects individually, in all 20 subjects, there is a saving in setting time with the favorable ratios. In 19 of the 20 subjects, there is likewise a decrease in variability, and in 18 of the 20 subjects a decrease in mis-settings.

When the data are analyzed according to the nine target positions, at each target position there is a saving in setting time, a decrease in variability, and a reduction in mis-settings with the favorable ratios.

*Different Lever Lengths with Favorable L/P Ratios.* The aim of the next set of experiments was two-fold: to determine whether lever-length as such was significant within the favorable L/P ratios, and to see whether there

Table 1  
Each Value is the Mean of 3,600 Settings  
(20 Subjects X 9 Target Positions  
X 20 Settings)

L/P Ratio	Lever Length	Mean Setting Time	Mean Variability (rms of $\sigma$ 's)*	Mis-settings
1.0	12"	2.58 sec.	0.80 sec.	7.4%
1.0	18"	2.48 sec.	0.70 sec.	6.6%
1.4	24"	2.20 sec.	0.54 sec.	4.8%
1.6	12"	2.23 sec.	0.58 sec.	5.0%
1.7	30"	2.18 sec.	0.54 sec.	3.9%
2.0	18"	2.02 sec.	0.46 sec.	4.3%
2.0	24"	2.02 sec.	0.44 sec.	4.8%
2.5	30"	1.99 sec.	0.40 sec.	3.7%
2.7	24"	1.93 sec.	0.37 sec.	3.3%
3.1	24"	1.92 sec.	0.35 sec.	4.0%
3.4	30"	1.94 sec.	0.33 sec.	3.2%
3.9	30"	1.95 sec.	0.34 sec.	3.5%

\* rms is the square root of the mean of the squares of the standard deviations, i.e.,

$$\sqrt{\frac{\sigma_1^2 + \sigma_2^2 + \dots + \sigma_n^2}{n}}$$



was any indication of an optimal ratio within the favorable region. Accordingly, lever lengths of 12, 18, 24, and 30 inches were employed with apparatus settings to give L/P ratios of 2.0, 2.5, and 3.0 (except that it was not possible to reach an L/P ratio of 3.0 with the 12" lever in the present apparatus). Target positions were restricted to the four on the three-inch diameter and the six on the seven-inch diameter. Each of 19 subjects made 20 settings at each of 10 positions using each of the 11 lever-ratio combinations.

Table 2 shows the findings. It is evident that lever-length as such plays little or no part in the outcome. However, the L/P ratio of 2.5 is slightly superior to 2.0 in setting time, variability, and mis-settings, and inferior to 3.0 only in mis-settings. For convenience the L/P ratio of 2.5, being lower, will be considered optimal.

*Starting Positions.* Up to this point the starting position of the cursor was always at the bottom of the simulated scope face. The question was raised whether this was the best starting position. In the next series of experiments, five starting positions were used: top, bottom, right, left, and center of the ten-inch diameter circle on which the outer seven inserts were located. Each of 12 subjects made 20 settings at each of 17 target posi-

Table 2

Lever Length and Optimal L/P Ratio

Note: Each value is the mean of 3,800 settings (19 subjects  $\times$  10 target positions  $\times$  20 settings).

L/P Ratio	Lever Length	Mean Setting Time	Mean Variability (rms of $\sigma$ 's)	Mis-settings
2.0	12"	1.71 sec.	0.32 sec.	2.2%
	18"	1.66 sec.	0.33 sec.	2.4%
	24"	1.64 sec.	0.31 sec.	2.4%
	30"	1.72 sec.	0.35 sec.	2.2%
2.5	12"	1.63 sec.	0.28 sec.	2.2%
	18"	1.59 sec.	0.26 sec.	2.1%
	24"	1.59 sec.	0.27 sec.	2.3%
	30"	1.63 sec.	0.30 sec.	2.0%
3.0	12"	—	—	—
	18"	1.60 sec.	0.26 sec.	1.4%
	24"	1.57 sec.	0.25 sec.	1.7%
	30"	1.59 sec.	0.25 sec.	1.4%

Table 3

Influence of Starting Position on Performance

L/P ratio = 2.5    Lever = 24"

Note: Each value is the mean of 4,080 settings (12 subjects  $\times$  17 target positions  $\times$  20 settings).

Starting Position	Mean Setting Time	Mean Variability (rms of $\sigma$ 's)	Mis-settings
Top	2.08 sec.	0.38 sec.	2.2%
Bottom	1.99 sec.	0.39 sec.	2.9%
Right	2.01 sec.	0.38 sec.	1.9%
Left	2.01 sec.	0.40 sec.	2.6%
Center	1.88 sec.	0.36 sec.	3.2%

tions (including the seven on the ten-inch diameter), using the L/P ratio of 2.5, and starting from each of the five positions. Average travel distance was thus the same for all starting positions.

Table 3 shows the results. In terms of mean setting time and variability the center position is slightly superior. On the other hand, in percentage of mis-settings the center starting position is the worst of the five.

In another analysis of the same data, the *best* starting position for each subject was determined in terms of each of the three criteria. In setting time, the center position is best for eight out of twelve subjects. In variability, the center position is best for five subjects. But in mis-settings no position stands out as being best. In overall view, it seems that starting position is relatively unimportant.

*Reversed Front-Back Operation.* In normal operation, the cursor moved upward when the joy-stick was pushed away from the subject and downward when the joy-stick was pulled toward the subject. A question was raised concerning the effect on the optimal L/P ratio if this operation was reversed so that the cursor moved upward when the joy-stick was pulled toward the subject and vice versa.

Each of 17 subjects made 10 settings at each of the 10 inner target positions with each of five ratios (Trials 1-10), using the normal direction of operation. He then made 40 settings at each of the 10 target positions with each of five ratios (Trials 11-20, 21-

30, 31-40, and 41-50), using the reversed direction of operation. Finally he made another 10 settings at each of the 10 target positions with each of the five ratios (Trials 51-60).

Table 4 shows by blocks of 10 trials the results in mean setting time, mean variability, and mis-settings. Two points can be noted: First, by the end of 40 trials with reversed operation (comprising 2,000 individual settings per subject) performance with reversed operation approached performance with direct operation, indicating that the subjects learned to handle what they all called an unnatural relationship of joy-stick and cursor move-

ment. Second, for both conditions, an L/P ratio of 2.5 is the lowest that can be called optimal.

*Subject's Switch.* In all the studies just described, the subject's switch was held in the hand that was not operating the joy-stick. A question was raised as to whether other types of switching would affect the performance. Two other types of switches were added: A push-button was located at the top of the upper end of the joy-stick, operating with very light pressure. A foot-pedal, with enough spring resistance to bear the weight of the subject's foot, was placed at a convenient position on the floor.

Table 4  
Performance with Reversed Direction of Operation of Joy-stick and Cursor  
Note: Each value is the mean of 1,700 settings.  
(17 subjects  $\times$  10 target positions  $\times$  10 settings)  
Lever Length 24"

		Mean Setting Time (seconds)				
Trial Nos.	Operation	L/P Ratios				
		1.4	1.9	2.2	2.5	3.0
1-10	(Direct)	(1.81)	(1.65)	(1.68)	(1.66)	(1.64)
11-20	Reversed	2.16	2.05	2.09	2.02	2.03
21-30	Reversed	2.03	1.89	1.86	1.91	1.88
31-40	Reversed	1.95	1.82	1.80	1.78	1.80
41-50	Reversed	1.84	1.74	1.72	1.69	1.72
51-60	(Direct)	(1.72)	(1.58)	(1.58)	(1.56)	(1.52)
		Mean Variability (rms of $\sigma$ 's in sec.)				
Trial Nos.	Operation	L/P Ratios				
		1.4	1.9	2.2	2.5	3.0
1-10	(Direct)	(0.38)	(0.29)	(0.30)	(0.28)	(0.25)
11-20	Reversed	0.49	0.44	0.48	0.43	0.41
21-30	Reversed	0.42	0.34	0.34	0.35	0.33
31-40	Reversed	0.38	0.33	0.32	0.31	0.32
41-50	Reversed	0.36	0.30	0.30	0.31	0.29
51-60	(Direct)	(0.32)	(0.25)	(0.25)	(0.25)	(0.23)
		Mis-settings (percentage)				
Trial Nos.	Operation	L/P Ratios				
		1.4	1.9	2.2	2.5	3.0
1-10	(Direct)	(11.5%)	(7.9%)	(7.2%)	(6.7%)	(6.6%)
11-20	Reversed	10.5	7.2	8.1	7.1	6.1
21-30	Reversed	8.6	6.9	7.6	4.8	7.0
31-40	Reversed	6.9	5.1	5.4	5.0	5.5
41-50	Reversed	6.9	6.2	5.4	4.9	5.4
51-60	(Direct)	(6.9)	(5.8)	(4.7)	(4.8)	(3.8)

Table 5

Influence of Type of Switch on Performance

Note: Each figure is the mean of 5,100 settings (10 subjects  $\times$  17 positions  $\times$  30 settings).

Ratio 2.5, Lever Length 24"

Type of Switch	Mean Setting Time	Mean Variability (rms of $\sigma$ 's)	Mis-settings
Other hand	1.46 sec.	0.19 sec.	8.9%
Joy-stick tip	1.47 sec.	0.19 sec.	10.1%
Foot pedal	1.47 sec.	0.19 sec.	8.2%

Each of 10 subjects made 30 settings at each of 17 target positions with each of the three types of switches. A 24" lever and an L/P ratio of 2.5 were employed throughout.

Table 5 shows the results. It is evident that all three types of switches are about equal in terms of mean setting time, mean variability, and mis-settings. Apparently any one of these three types of switches, whichever is most convenient, can be used without affecting performance.

#### Summary

A series of experiments was performed to determine the significance of certain variables in the use of a joy-stick to make settings in two dimensions on a simulated scope face to a relatively coarse tolerance.

The most significant factor turns out to be the ratio between the movement of the joy-stick tip and the movement of the cursor. The lowest ratio that can be considered optimal is about two-and-a-half. That is, the tip of the joy-stick should move two-and-a-half times as fast as the cursor.

Other variables proved to be relatively unimportant. Joy-stick lengths of 12", 18", 24", and 30" are equally effective. Starting position (top, bottom, right, left, or center of the scope) makes little difference in the overall results. Reversed operation (cursor moving down when stick is pushed away from the operator) is slower but the optimal ratio is the same. Finally, results are not affected by the position of the subject's switch, whether it is operated by the hand not holding the joy-stick, by a foot-pedal, or by the same hand that moves the joy-stick.

It should be emphasized that these results were obtained in a situation where the movement of the joy-stick is translated directly into movement of the cursor. The present type of apparatus does not permit making tests of a similar nature with joy-stick controls where the movement of the pointer is determined by pressure rather than by extent of movement of the joy-stick.

*Received November 27, 1953.*



## Figure and Ground in a Two Dimensional Display \*

R. C. Browne

*The Nuffield Department of Industrial Health, University of Durham,  
King's College, Newcastle upon Tyne*

In an aircraft the subjective feelings of passenger or pilot are little guide to the attitude which the machine assumes, and they are even less guide when it is flying in the dark or in cloud when there is nothing to which external reference can be made. An indicator (Figure 1) was therefore developed to provide the pilot with a visible display of how the attitude of his aircraft varies in relation to an artificial horizon which is gyroscopically stabilized. It shows in two dimensions whether the aircraft is climbing, diving, or banking and also, on a scale, the amount of bank, in degrees from the horizontal. This display provides a "figure and ground" problem in that it is the horizon which apparently moves and not the aircraft. This does not, of course, accord with the facts, although it does with the appearance of the horizon as seen from the aircraft. Because of this, it was thought likely that air pilots often made wrong control movements and so increased the departure of their machine from the straight and level attitude. To meet these objections a new display was designed in which a diagrammatic aircraft moved in reference to a stationary horizon. The problem in its crude *ad hoc* form was, therefore, to decide which of these two displays was the more suitable.

An initial examination of the two displays showed that they differed in three respects:

1. In the old method (Figure 1D) the "figure" or miniature aircraft is stationary, whereas in the new (Figure 1A) it moves against a "ground" composed of a horizon which is still.

2. The old display is provided with a scale and pointer which shows how many degrees the aircraft is banking to one side or the other, but in the reversed sense.

\* Acknowledgments are due to the Medical Directorate, British Royal Air Force, for permission to publish this paper, and to Mr. H. Campbell, B.A., F.S.S. for statistical advice.

3. The old instrument is less heavily "damped" than the new; in other words, the new display takes rather longer to come to rest after a given deflection.

### Method

The classical method of studying a display problem is with a tachistoscope. But, on the other hand, where machinery is controlled in response to alterations in an indicator (as in the present study) and some movement in a control system has to be made, it is perhaps better to assess the different displays in a comparable way by requiring the experimental subjects to make control movements in response to changes in them, and to measure the speed and accuracy with which they do so.

A standard instrument flying trainer was, therefore, used as the machine to be controlled, and it was fitted with a recording apparatus which integrated the speed and accuracy with which deflections from the straight and level attitude were corrected. It gave a numerical score every two minutes. The test lasted for eleven minutes which allowed time for four such scores to be made and noted. The attitude of the machine in the test was made to change quickly in a cyclical fashion which repeated itself every eighteen seconds. The task before the subjects of the experiment was to correct the changes in attitude which were conveyed by one or other of the two indicators. The hood of the trainer was shut, so that no fixed external reference point could be seen, and it was so arranged that there were no turning movements. A number of cadets chosen at random from a large group who had already been selected to be air pilots and who were, therefore, quite homogeneous, were the subjects of the experiment. But they were at a stage in training when they had had no experience in atti-

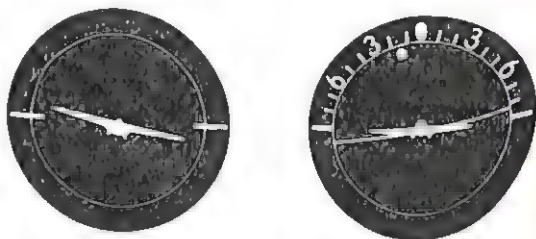


FIG. 1. The two displays. The New (A) is on the left and the Old (D) on the right.

Table 1

The Numbers in Each Group of Subjects and the Order in Which They Were Tested on the Various Displays

Display Order of Test	Subjects	
	No. of	Group Letter
1. D	20	a
A	20	b
2. D	10	c
A	10	d
A	10	c
D	10	d
3. C	20	e
4. B	20	f
Trained Air Pilots		
5. D	10	g
A	10	h
A	10	g
D	10	h

tude display indicators. In this way, bias due to familiarity with either display was avoided. Every man received a comparable explanation, and was allowed to practice until he could just do the test without damaging the apparatus. The test was kept short—eleven minutes—to avoid fatigue and fluctuating levels of attention.

The experiment was divided into five parts as shown in Table 1.

1. Two groups of 20 subjects (a) and (b) were chosen. Group (a) was tested on the old indicator (D) and Group (b) on the new indicator (A).

2. Twenty new subjects were chosen and divided into two groups of ten (c) and (d). Group (c) was first tested on the old indicator (D) and then on the new indicator (A). Group (d) carried out the same two tests in the reverse order.

3. The old indicator (D) was partially covered with black paper to make it comparable to (A) in every respect except the figure and ground relation and the degree of damping. A new group of 20 subjects (e) was tested on this display (C).

4. The display (C) was further modified so that the damping was comparable to (A) and another group of 20 subjects (f) was tested on this new display (B). (B) now resembled (A) in every respect except the figure and ground relation.

5. Two groups of ten experienced pilots (g) and (h) who had trained on the old display (D) were chosen. Group (g) was first tested on the old indicator (D) and afterwards on the new indicator (A). Group (h) carried out the same

two tests in the reverse order. For this experiment the damping was made comparable on the two displays.

## Results

Figure 2 plots the means of the four scores for every subject for the roll or side-to-side movements, and Figure 3 shows similar scores for the pitch or fore and aft movements. In parts 1 and 2 of the experiment 30 subjects (groups a and c, Table 1) had their first test on the old display (D) and another comparable 30 (groups b and d) on the new (A). It was found that 5.0 fewer errors were made in roll on (A) than on (D) which seems unlikely to be due to chance, since  $t = 2.75$  and  $P = 1$  in 100. In pitch the difference is smaller (1.5 errors) as, indeed, were the disturbances in attitude to be corrected. But here too, fewer errors are made with the new display (A). Taken alone, this might be due to chance, but the difference is in the same direction as the difference in roll which lends, therefore, a certain weight to it. The instruction times needed before the subjects were fit to start the test and their preferences for the two displays, are shown in Table 2. These two criteria were measured for 20 of the 30 men in each group, and show that significantly less instruction time was needed in the case of the new display (17 compared to 23.5 minutes) which was also subjectively preferred by between six and seven times as many of the men (33 compared to 5) who were tested upon it.

The results with displays (C) and (B) in roll with fresh groups of subjects fall into intermediate positions between the other two

Table 2

The Length of Instruction Time Needed Before the Test Could be Started and the Subjective Preferences for the Two Displays

Instruction Time Minutes	Display		
	Old	New	Neither
Mean	23.5	17.0	
Difference	6.5 $\pm$ 1.84		
Number of Men with Preference for	5	33	2

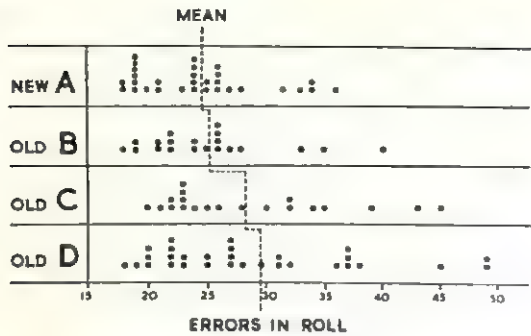


FIG. 2. Each subject's individual scores in roll and the means for the groups.

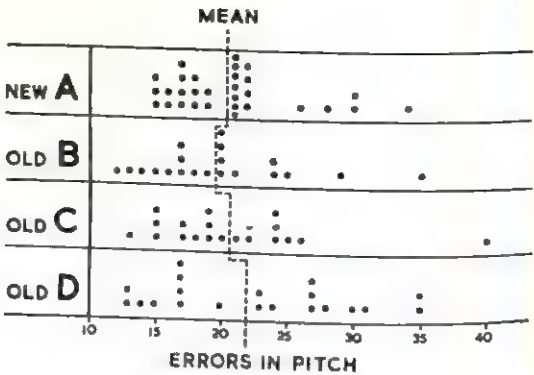


FIG. 3. Each subject's individual scores in pitch and the means for the groups.

as, indeed, did the displays themselves. In pitch (Figure 3), however, there was little to choose between the results given by the different designs.

The relations between the means and standard deviations of these figures on the four types of display are of some interest, and they are shown in Table 3 and Figure 4. In the roll dimension, as the display becomes easier to interpret through the sequence D, C, B, and A, and errors fall from 29.7 to 24.7, so the scatter between subjects falls also from 8.5 to 5.1. But the scatter within a given subject's performance remains much more constant at between 3.3 and 4.3 errors. The figures for pitch demonstrate the same trend less markedly. As the test becomes harder it magnifies the individual differences, but it does not appear to make the performance of a single man more erratic.

Ten subjects (groups c and d, Table 1) were tested upon each of the two displays

after previous experience with the other, to satisfy the desire for a "double" experimental design and to investigate the question of transfer. Errors were again fewer with display (A) when first test was compared with first, and second with second (Table 4). There is positive transfer from test to test (Table 5), whichever of the two came first, but with a difference in degree. Previous experience with display (D) stood the subjects in better stead than previous experience with (A) and this was more marked in the roll dimension in which the disturbances to be corrected were the greater. The positive transfer from (D) to (A) in roll was four times as great as in the reverse direction, and in pitch it was in the ratio of 1.6:1. This is to be expected if the new display (A) is easier to read than the old (D), and it makes the point that in this type of problem it is unsafe, in designing the experiment, to as-

Table 3  
The Relationship of the Means to the Standard Deviations Between and Within Subjects

Display	Dimension					
	Roll			Pitch		
	Mean Errors	Standard Deviation		Mean Errors	Standard Deviation	
		Between Men	Within Man		Between Men	Within Man
New A	24.7	5.1	3.9			
Old B	25.2	5.6	3.5	20.6	6.1	3.2
Old C	28.4	7.6	3.3	19.8	5.6	3.1
Old D	29.7	8.5	4.3	20.7	5.8	3.7
				22.1	6.7	3.2



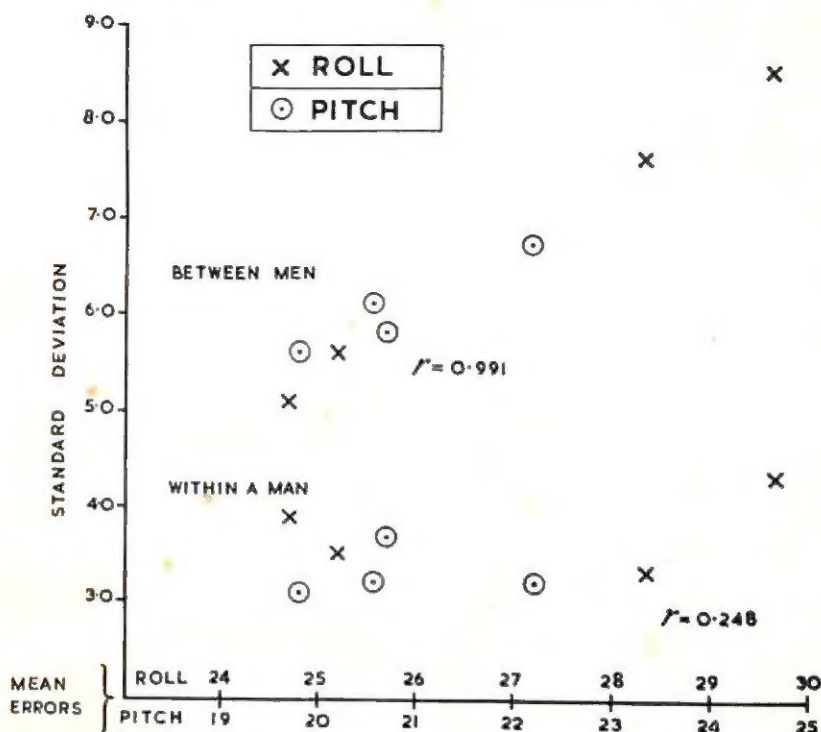


FIG. 4. The relationship of the means to the standard deviations between and within subjects.

sume an equal amount of transfer in both directions.

In the final part of the experiment, two groups (g) and (h) of ten air pilots who had about 300 hours experience with the old display were tested on both displays in alternate order. Both groups had more men who preferred the new design (Table 6), but the differences in performance, while they slightly favored this design in pitch, were small and might have been due to chance. It is, perhaps, noteworthy that with so much previous

Table 5

Positive but Unequal Transfer between the Displays

Transfer from	Pitch	Roll
(D) to (A)	+6.5	+12.0
(A) to (D)	+4.0	+ 3.1

Table 6

The Experience, Preference and Performance of the Trained Subjects

Group of 10 Subjects		g	h
Experience (hrs)		318	292
No. with preference for:—	New (A)	7	7
	Old (D)	2	2
	Neither	1	1
Errors in:—	Pitch:	New (A)	9.4
		Old (D)	10.3
	Roll:	New (A)	11.8
		Old (D)	11.5

Table 4

Fewer Errors with One Display after Experience with the Other

Subjects' Group Letter	Display				
	Old (D)		Test Sequence	New (A)	
	Pitch	Roll		Pitch	Roll
c	20.7	29.2	→	14.2	17.2
d	16.3	20.8	←	20.3	23.9

practice with the old design (D) they were not worse than they were when tested with the new (A).

### Discussion

Craik (2) has pointed out that when an air pilot has a good unobstructed view of the external world, as in day flying, the aircraft can be considered to be an extension of his body which moves with him and which he can orientate in direct reference to the background of the external horizon. But if, on the other hand, the view is obstructed or absent, as at night or in cloud, the interior of the machine itself becomes the external environment or background, and the pilot is then faced with the paradoxical fact that he and his background remain relatively fixed however he manipulates the controls. Craik, therefore, suggested a much larger representation of the moving horizon, as in the old attitude indicator (D), as a way out of this situation. However, this may not entirely ensure that emergence of figure from ground, which forms the essential part of perception in this kind of display (Vernon, 4). Where the contrast between figure and ground is small the results of the present study suggest that it does not matter which of the two is moving and which fixed. But from the point of view of the immediate *ad hoc* problem of whether the aircraft or the horizon should move it can be argued *a priori* that it should be the aircraft. According to Rubin's classification (Woodworth, 5) the aircraft has the characteristics of the "figure" rather than of the "ground," because: (1) it has form while the horizon bar is relatively formless; (2) the aircraft tends to appear in front, the horizon behind; and (3) the aircraft is more impressive and "more apt to suggest meaning."

In the design of any experiment of this kind to investigate two different displays, two difficulties have to be considered: (1) the comparability of the groups of subjects used; and (2) performance transfer, either positive or negative, from one test to the other.

If one group of subjects tested on one dis-

play is compared with another group on a second display, any observed difference may, on the face of it, be either an intrinsic function of the group or of the display. The use of large groups to which the subjects are allocated at random after careful matching by some analogous test, in theory, helps to ensure their equivalence. But, in practice, the matching has to be demonstrably analogous to the problem in hand. It is not safe merely to assume or to assert this; neither is it easy usually to demonstrate this analogy, which at best means a lengthy piece of experimental work. The alternative design in which one half of the subjects is tested on one display and then on the other, and the other half of the subjects vice versa, can equally well be criticized on the ground that the second tests are only comparable if the amount of transfer is the same in both directions, which may well be unlikely (as in the present study), if one display is easier to perceive than the other. The conclusion seems to be that the experimental design must be arbitrary to a certain extent, and that the most secure design is, perhaps, a combination of both these methods.

In an experiment which is generally comparable to that described here, Loucks (3) showed that a display having a reversed sense to that of the "old" (D) indicator described in this paper produced a greater speed and accuracy of response than did one similar to the old indicator itself. He was also using subjects with no previous experience who appeared to identify themselves with the moving component of the display irrespective of its appearance. He suggests that it would be even better if the moving component were drawn in the shape of a small aircraft. But an experiment with this type of display was not, in fact, tried, and the present study suggests that this change might have made little difference. However, the numbers of subjects used in it were relatively small and the subject must still be considered open. It seems clear that in order to alter the ease of perception, figure and ground must contrast in qualities other than mere relative movement, which alone seems unimportant.



## Summary

1. The speed and accuracy of human response to two displays which give information in two dimensions has been compared. Comparisons of the instruction times needed before the test could be started, and of the preferences of the subjects, have also been made.

2. The two displays differed in respect of: (i) the relation of figure and ground and their relative movement; (ii) the damping of the oscillations after a given displacement; and (iii) their relative complication.

3. The speed and accuracy of response was greater with the more simple display which had the heavier damping. This display also needed a shorter instruction time and was preferred more often by the subjects. Within the limits of the experimental design employed the pure figure and ground relation alone appeared to play little part in perception.

4. The individual differences between subjects increased as perception became more

difficult. But the differences between different samples of a single subject's performance remained constant.

5. In an experimental design learning transfer between different tests must not be considered to be the same. Neither is matching groups of subjects on an assumed analogous test experimentally safe.

Received November 24, 1953.

## References

1. Browne, R. C. *Comparative trial of two attitude indicators*. Royal Air Force Flying Personnel Research Committee Reports Nos. 611 and 611a, Feb. and April, 1945.
2. Craik, K. J. W. *Figure and ground in control of aircraft*. (Unpublished), 1944.
3. Loucks, R. B. *An experimental evaluation of the interpretability of various types of aircraft attitude indicators*. Psychological Research on Equipment Design, Report No. 19, p. 111. Washington: U. S. Government Printing Office, 1947.
4. Vernon, M. D. *Visual perception*. Cambridge University Press, 1937.
5. Woodworth, R. S. *Experimental psychology*. London: Methuen, 1950.





## New Books, Monographs, and Pamphlets

Books, monographs, and pamphlets for listing and possible review should be sent to Dr. John G. Darley, Editor-elect, Graduate School, University of Minnesota, Minneapolis 14, Minnesota.

- A study of personnel practices for college and university office and clerical workers.* Wilbur Donald Albright. Champaign, Ill.: College and University Personnel Association, University of Illinois, 1954. Pp. 131. \$2.00.
- Counseling in the Y.M.C.A.* Seth Arsenian and Francis W. McKenzie. New York: Association Press, 1954. Pp. 126. \$2.00.
- The social psychology of industry.* J. A. C. Brown. Baltimore: Penguin Books, Inc., 1954. Pp. 309. \$.65, Paperback.
- Remaking men.* Paul Campbell and Peter Howard. New York: Arrowhead Books, Inc., 1954. Pp. 126. \$1.50.
- Readings in general psychology.* Lester D. Crow and Alice Crow. New York: Barnes & Noble, Inc., 1954. Pp. 437. \$1.75, Paperback.
- Dark destiny.* Edgar E. Daniels. New York: Vantage Press, Inc., 1954. Pp. 172. \$3.00.
- Adjusting to a competitive economy—the human problem.* M. J. Doohar, Editor. New York: American Management Association, 1954. Pp. 48. \$1.25.
- Perceptualistic theory of knowledge.* Peter Fireman. New York: Philosophical Library, 1954. Pp. 50. \$2.75.
- Psychometric methods.* Second Edition. J. P. Guilford. New York: McGraw-Hill Book Company, Inc., 1954. Pp. 597. \$8.50.
- Nebraska symposium on motivation.* Marshall R. Jones, Editor. Lincoln: University of Nebraska Press, 1954. Pp. 322. \$3.50, Cloth; \$3.00, Paperback.
- Conflict and mood.* Patricia Kendall. Glencoe, Ill.: The Free Press, 1954. Pp. 182. \$3.50.
- Psychomotor aspects of mental disease: An experimental study.* H. E. King. Cambridge, Mass.: Harvard University Press, 1954. Pp. 185. \$3.50.
- Industrial conflict.* Arthur Kornhauser, Robert Dubin, and Arthur M. Ross, Editors. New York: McGraw-Hill Book Company, Inc., 1954. Pp. 551. \$6.00.
- Mathematical thinking in the social sciences.* Paul F. Lazarsfeld, Editor. Glencoe, Ill.: The Free Press, 1954. Pp. 444. \$10.00.
- The sexual nature of man and its management.* Clarence Leuba. New York: Doubleday and Company, Inc., 1954. Pp. 40. \$.85.
- Effective leadership in human relations.* Henry Clay Lindgren. New York: Hermitage House, Inc., 1954. Pp. 287. \$3.50.
- A psychological approach to accidents.* Norman Roberts Lykes. New York: Vantage Press, Inc., 1954. Pp. 138. \$2.95.
- The encyclopedia of child care and guidance.* Sidonie Matsner Gruenberg, Editor. New York: Doubleday and Company, Inc., 1954. Pp. 1016. \$7.50.
- Studying and learning.* Max Meenes. New York: Doubleday and Company, Inc., 1954. Pp. 68. \$.95.
- School and child: A case history.* Cecil V. Millard. East Lansing: Michigan State College Press, 1954. Pp. 221. \$3.75.
- Aspects of readability in the social studies.* Eleanor M. Peterson. New York: Bureau of Publications, Teachers College, Columbia University, 1954. Pp. 118. \$3.50.
- Psychotherapy and personality change.* Carl R. Rogers and Rosalind F. Dymond, Editors. Chicago: University of Chicago Press, 1954. Pp. 445. \$6.00.
- Basic concepts in vocational guidance.* Herbert Sanderson. New York: McGraw-Hill Book Company, Inc., 1954. Pp. 338. \$4.50.
- The laws of life.* Adrian Waldo Sasha. San Francisco: Living Knowledge Foundation, 1954. Pp. 224. \$5.00.
- The real enjoyment of living.* Hyman Judah Schachtel. New York: E. P. Dutton & Co., Inc., 1954. Pp. 192. \$2.75.
- The mind and the universe.* Charles R. Smith. New York: The William-Frederick Press, 1954. Pp. 173. \$3.50.
- Decision-making as an approach to the study of international politics.* Richard C. Snyder, H. V. Bruck, and Burton Sapin. Princeton, N. J.: Organizational Behavior Section, Princeton University, 1954. Pp. 120.
- An inventory of social and economic research in health.* Frederick R. Strunk, Editor. New York: Health Information Foundation, 1954. Pp. 180.
- The prediction of student-teaching success from personality inventories.* Fred T. Tyler. Berkeley: University of California Press, 1954. Pp. 31. \$12.50.
- Psychology as a profession.* Robert I. Watson. New York: Doubleday and Company, Inc., 1954. Pp. 65. \$.95.
- Human engineering guide for equipment designers.* Wesley E. Woodson. Berkeley: University of California Press, 1954. Pp. 259. \$3.50.
- Lotteries-for-housing.* Martin Zethfield. New York: The William-Frederick Press, 1954. Pp. 26. \$1.00.

### Correction

In the New Books section of the August issue the *Journal of Applied Psychology* on page 282 the price of Anastasi's *Psychological Testing* was listed as \$4.25. The actual list price is \$6.75.